# شبکه‌های عصبی در

# SPSS

**مؤلف: Marija Norusis**

**(SPSS Company)**

**مترجم: امیررضا فتی‌پورجلیلیان**

**مازیار نجبا**

# فهرست مطالب

# پیشگفتار

در جهان امروز به علت پیشرفت تکنولوژی و پیچیده‌تر شدن مسائل، استفاده از روش‌های نوین، جایگزین بسیاری از روش‌های سنتی شده که دیگر قادر به تخمین درستی از وضعیت موجود نمی‌باشند. همچنین پیچیده شدن فرایندها منجر به مشکلاتی مانند غیرخطی شدن رابطه‌ی پارامترهای فرایند شده که روش‌های پیشین قادر به انجام و یا تصمیم‌گیری در مورد آنها نیستند، از این‌رو روش‌های جدیدی از قبیل شبکه‌های عصبی جهت تحلیل این فرایندها پدید آمده است.

به دلیل پیچیده بودن محاسبات شبکه عصبی استفاده از نرم‌افزارهای کامپیوتری توسط کاربران ناگزیر می‌نماید. در این میان نرم‌افزارهای متعددی به محاسبه شبکه‌های عصبی می‌پردازند. نرم‌افزار SPSS نسبت به سایر نرم‌افزارها دارای مزیت‌هایی بوده که مهمترین آن سهولت استفاده از آن می‌باشد.

از این‌رو برآن شدیم تا راهنمای مناسبی جهت استفاده از این نرم‌افزار مهیا کنیم که کتاب حاضر حاصل این احساس نیاز می‌باشد. توجه داشته باشید که این آموزش براساس مثال‌های کاربردی صورت می‌پذیرد که تأثیر مفیدی بر آموزش نرم‌افزار دارد.

با امید به اینکه از این کتاب بهره کافی را داشته باشید.

*امیررضا فتی‌پورجلیلیان*
*مازیار نجبا*

# فصل اول

# مقدمه

## شبکه عصبی چیست؟

اصطلاح شبکه عصبی به خانواده‌ای از مدل‌ها اشاره می‌کند که با یک فضای بزرگ پارامتری و ساختار منعطف مشخص شده و از روی مطالعات مغزی الهام گرفته شده است. با بزرگ شدن این خانواده، اکثر مدل‌های جدید برای کاربردهای غیربیولوژیکی طراحی شده‌اند، گرچه اکثر اصطلاحات فنی مرتبط، ریشه بیولوژیک این کلمات را نشان می‌دهند. تعاریف تخصصی شبکه‌های عصبی، از آنجایی که این شبکه‌ها درگستره بزرگی از کاربردها مفید و کاربردی می‌باشند، متنوع است و به این دلیل که هیچ تعریف جامعی که بتواند تمام مدل‌های موجود در این خانواده را پوشش دهد وجود ندارد، در حال حاضر از تعریف زیر استفاده می‌شود (هایکین، ۱۹۹۸)

شبکه عصبی یک "**پردازنده توزیع شده موازی**[1]" است که میل طبیعی برای ذخیره دانش تجربی و قابل استفاده کردن آن دارد. از دو جهت به مغز شباهت دارد.

- دانش از طریق یک فرایند یادگیری توسط شبکه کسب می‌شود.
- قدرت ارتباط بین نرونی که به‌عنوان وزن‌ها سیناپسی[2] شناخته می‌شود، برای ذخیره دانش مورد استفاده قرار می‌گیرد.

برای تمیز شبکه عصبی از روش‌های آماری مرسوم، چیزی که گفته نشده است به اندازه متن واقعی تعریف مهم می‌باشد. برای مثال، مدل مرسوم رگرسیون خطی می‌تواند با روش حداقل مربعات اطلاعات را جمع‌آوری کرده و آنها را به‌صورت ضریب رگرسیون ذخیره کند. از این منظر این روش یک شبکه عصبی است. در واقع، می‌توان این طور استدلال کرد که رگرسیون خطی یک حالت خاص از شبکه‌های عصبی مشخص است. با این تفاوت که، رگرسیون خطی دارای یک ساختار مدل نامنعطف و مجموعه فرضیاتی است که قبل از یادگیری اطلاعات اعمال می‌شوند. تعریف بالا نیاز به ساختار مدل و فرضیات را حداقل می‌کند. بنابراین یک شبکه عصبی می‌تواند بازه وسیعی از مدل‌های آماری را بدون نیاز به فرض رابطه مشخص بین متغیرهای وابسته[3] و مستقل[4] ، تخمین بزند. در عوض، نوع ارتباط حین فرایند یادگیری

1. parallel distributed processor
2. Synaptic Weights
3. Dependent variable
4. Independent variable

مشخص می‌شود. درصورتی‌که رابطه خطی بین متغیرهای مستقل و وابسته مناسب باشد، نتایج شبکه‌های عصبی باید به تخمین مدل رگرسیون خطی نزدیک باشند. اگر رابطه غیرخطی مناسب‌تر باشند، شبکه عصبی به‌صورت خودکار ساختار صحیح مدل را تخمین خواهد زد.

بهای این انعطاف‌پذیری، غیرقابل تفسیر بودن وزن‌های سیناپسی یک شبکه عصبی است. بنابراین چنانچه سعی در تشریح فرایندی دارید که به ایجاد روابط میان متغیرهای مستقل و وابسته می‌پردازد بهتر است که از مدل‌های آماری سنتی استفاده کنید. با این وجود چنانچه قابلیت تفسیر مدل برایمان حائز اهمیت نباشد می‌توان با استفاده از شبکه‌های عصبی سریعتر به نتایج دست یافت.

## مزیت‌های شبکه‌های عصبی

- شبکه عصبی، به دلیل پردازش‌های موازی، از سرعت پردازش بالایی برخوردار است.
- شبکه‌های عصبی توان بالقوه‌ای برای حل مسائلی دارند که شبیه‌سازی آن‌ها از طریق منطقی و یا سایر روش‌ها، مشکل و یا غیرممکن است.
- شبکه‌های عصبی همانند مغز انسان به‌طور پیوسته در حال یادگیری و انطباق با محیط هستند. به این معنی که اگر شبکه برای یک وضعیت خاص آموزش دید و تغییر کوچکی در شرایط محیطی آن رخ داد، می‌تواند با آموزش مختصر، برای شرایط جدید نیز کارآمد باشد.
- در شبکه عصبی، عدم عملکرد صحیح قسمتی از نرون‌ها، موجب از کار افتادگی کامل مغز نمی‌شود و امکان اتخاذ تصمیم صحیح نیز وجود دارد.
- این روش قادر است برای داده‌ها در شریط عدم اطمینان (اعم از آنکه فازی باشند و یا به‌طور ناقص و توام با دریافت noise دریافت شده باشند)، جواب منطقی ارائه دهد.

## محدودیت‌های شبکه عصبی

- شبکه‌های عصبی مصنوعی قادر به توضیح منطق و قاعده کار نیستند و اثبات درستی نتایج آنها بسیار دشوار است.
- محاسبات شبکه‌های عصبی معمولاً محتاج مقادیر زیادی برای آموزش مدل است.

- در حالت کلی، شبکه‌های عصبی برای برخی از مسائل کارایی ندارند. برای مثال برای حل مسائل و پردازش داده‌ها باروش مستدل مناسب نیستند.

## نرون عصبی انسان

در سیستم عصبی، نرون به‌عنوان اصلی‌ترین عنصر پردازش، شناخته شده‌اند. به‌طورکلی بدن انسان در حدود 100 تریلیون نورون وجود دارد که تمام آن‌ها از سه قسمت اصلی تشکیل شده‌اند: بدنه سلول[1]، دندریت‌ها[2] و آکسون[3] (شکل ۱-۱). همان‌طورکه در شکل مشخص است هر نرون دارای تعدادی دندریت و یک آکسون است. دندریت‌ها به‌عنوان مناطق دریافت‌کننده سیگنال‌های الکتریکی هستند و سیگنال‌های الکتریکی را از از آکسون نرون‌های دیگر به بدنه سلول می‌برند. بدنه سلول انرژی لازم را برای فعالیت نرون فراهم کرده و برروی سیگنال‌های ورودی عمل می‌کند (که با یک عمل جمع جبری ساده و مقایسه با یک سطح آستانه مدل می‌گردد.)

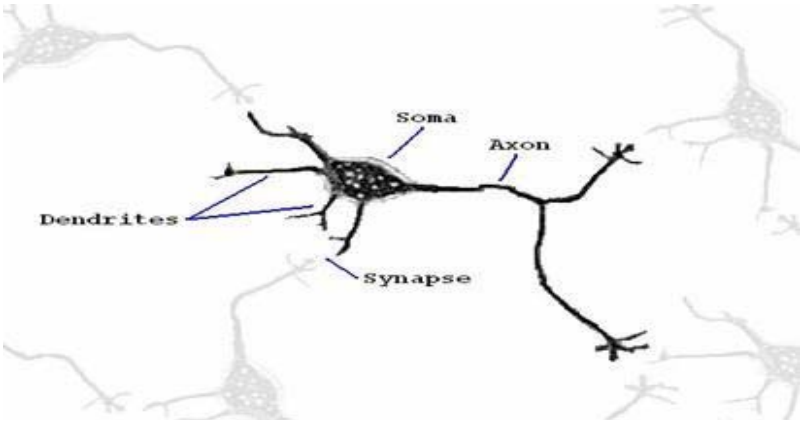
آکسون نیز سیگنال‌های الکتروشیمیایی را از بدنه سلول به دندریت سایر نورون‌ها منتقل می‌کند.

محل تلاقی یک آکسون از یک نرون به دندریت‌های سایر نرون‌ها را سیناپس[4] می‌نامند. سیناپس‌ها واحدهای کوچکی هستند که ارتباط بین نرون‌ها را برقرار می‌سازد.

زمانی‌که سیگنال های عصبی از آکسون سایر نرون‌ها به یک نرون می‌رسد، آن را تحریک می‌کند. نرون از هر یک از اتصالات ورودی خود یک ولتاژ کم را توسط سیگنال‌های عصبی، دریافت می‌کند و آن‌ها را با هم جمع می‌کند. اگر این مقدار به مقدار آستانه برسد نرون آتش می‌گیرد وبه آکسون خود یک ولتاژ خروجی ارسال می‌نماید و آکسون نیز با توجه به شدت آن، ممکن است یک سیگنال را توسط سیناپس، به دندریت نرون‌های دیگر بفرستد یا اینکه به دلیل ضعیف بودن آن، هیچگونه سیگنالی را عبور ندهد. و به همین ترتیب تمامی فعالیت‌های مغزی انسان انجام می‌شود.

---

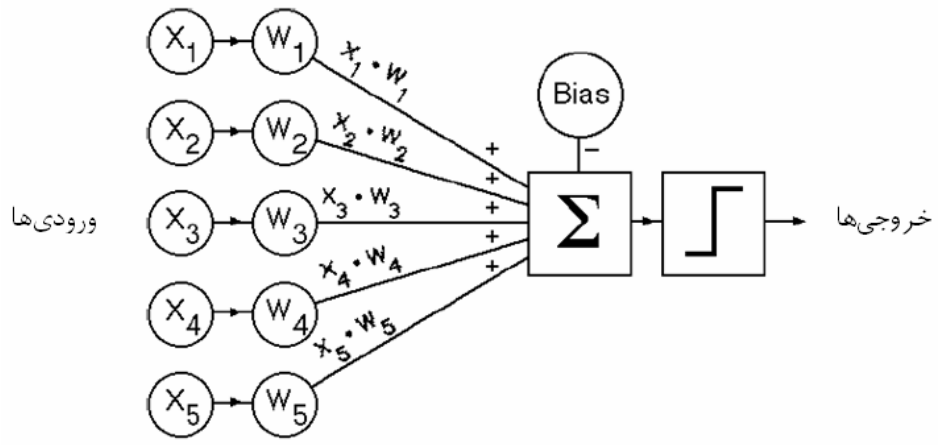
1. Cell body (Soma)
2. Dendrite
3. Axon
4. Synaptic




*شکل ۱-۱: ساختار سلول عصبی انسان*

با این دید اجمالی از نحوه عملکرد نرون، باید سیستمی طراحی شود که دارای تعدادی ورودی باشد که با توجه به اهمیت هر یک، آنها را با یکدیگر جمع ساده جبری نماید و توسط یک تابع موسوم به تابع تبدیل، آنها را به نورون‌های دیگرارسال نماید.

شکل ۱-۲ الگویی از یک واحد پردازش با توجه به نحوه عملکرد یک نرون ارائه می‌دهد. همان‌گونه که مشاهده می‌شود،آکسون را می‌توان به خروجی، وزن را به ولتاژ و ورودی‌ها را به دندریت‌ها تشبیه نمود.



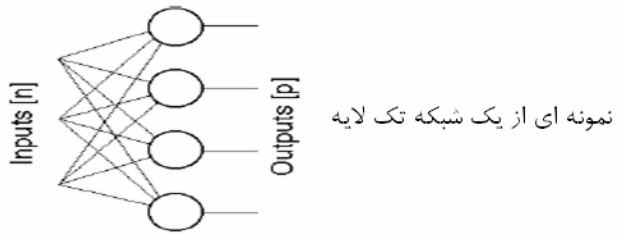

*شکل ۱-۲: ساختار یک نرون عصبی مصنوعی*

بنابراین اجزای یک شبکه عصبی عبارتند از:

- **ورودی‌ها**. ورودی‌ها می‌توانند خروجی سایر لایه‌ها بوده و یا آنکه به حالت خام در اولین لایه و به‌صورت‌های زیر باشد:
  - داده‌های عددی و رقمی
  - متون ادبی، فنی و...
  - تصویر و یا شکل
- **وزن‌ها**. میزان تأثیر ورودی $x_i$ بر خروجی مسئله را تا حدودی مشخص می‌کنند و در شبکه‌های چند نرونی نیز تابع جمع میزان سطح فعالیت نرون j در لایه‌های درونی را مشخص می‌سازد.
- **تابع تبدیل**. بدیهی است آن تابع جمع پاسخ مورد انتظار شبکه نیست. تابع تبدیل عضوی ضروری در شبکه‌های عصبی محسوب می‌گردد. انواع و اقسام متفاوتی از توابع تبدیل تبدیل وجود دارد که بنا به ماهیت مسئله کاربرد دارند. این تابع توسط طراح مسئله انتخاب می‌گردد و براساس انتخاب الگوریتم یادگیری، پارامترهای مسأله (وزن‌ها) تنظیم می‌گردد.
- منظور از خروجی، پاسخ مسئله است.

## شبکه‌های عصبی تک نرونه، تک لایه، چند لایه

معمولاً یک نرون با ورودی‌های زیاد، به تنهایی برای حل مسائل فنی- مهندسی کافی نیست. مثلاً برای مدل‌سازی نگاشت‌هایی که دو خروجی دارند ما احتیاج به دو نرون داریم که به‌صورت موازی عمل کنند، بنابراین یک لایه خواهیم داشت که از اجتماع چند نرون تشکیل شده است. در شکل ۱-۳ نمونه‌ای از یک شبکه‌ی تک لایه را نشان می‌دهد.

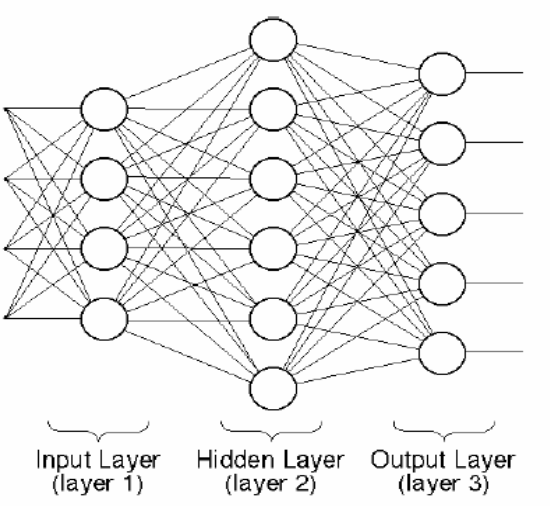



*شکل ۱-۳*

شبکه‌های تک لایه، توانایی پیاده‌سازی توابع غیرخطی را ندارند به همین دلیل از شبکه‌های که از چند لایه تشکیل شده‌اند استفاده می‌کنیم .این شبکه‌ها دارای توانایی بیشتری هستند.

در شبکه‌های عصبی چند لایه، یک لایه ورودی وجود دارد که اطلاعات را دریافت می‌کند، تعدادی لایه پنهان[1] وجود دارد که اطلاعات را از لایه‌های قبلی می‌گیرد (در اصل وجود لایه پنهان زمانی مفید است که تابع تبدیل غیرخطی باشد) و در نهایت یک لایه خروجی وجود دارد که نتیجه محاسبات به آنها رفته و خروجی آن، خروجی نهایی شبکه است.

نمونه‌ای از این شبکه‌ها را در شکل ۱-۴ می‌بینید:



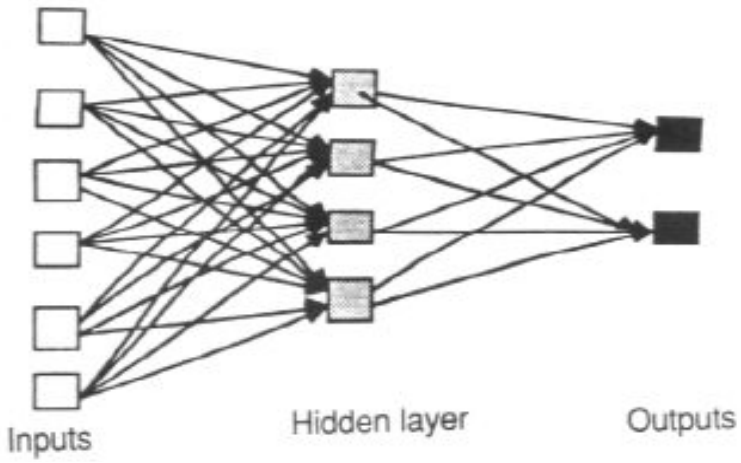

*شکل ۱-۴: نمونه‌ای از یک شبکه ۳ لایه*

## انواع شبکه‌ای عصبی مصنوعی از نظر برگشت‌پذیری

**شبکه‌های پیش‌خور**[2] . شبکه‌های پیش‌خور، شبکه‌هایی هستند که مسیر پاسخ در آنها، همواره رو به جلو پردازش می‌شود و به نرون‌های لایه‌های قبل بازنمی‌گردد. در این نوع شبکه‌ها به سیگنال‌ها اجازه می‌دهند تنها از مسیر یکطرفه عبور کنند، یعنی از ورودی تا خروجی. بنابراین بازخوردی وجود ندارد به این معنی که خروجی هر لایه تأثیری بر همان لایه ندارد. در بدن انسان نیز، پیام‌های عصبی به‌صورت یکطرفه حرکت می‌کنند. از دنریت به بدنه سلول و سپس به

1. Hidden Layer
2. Feed-Forward

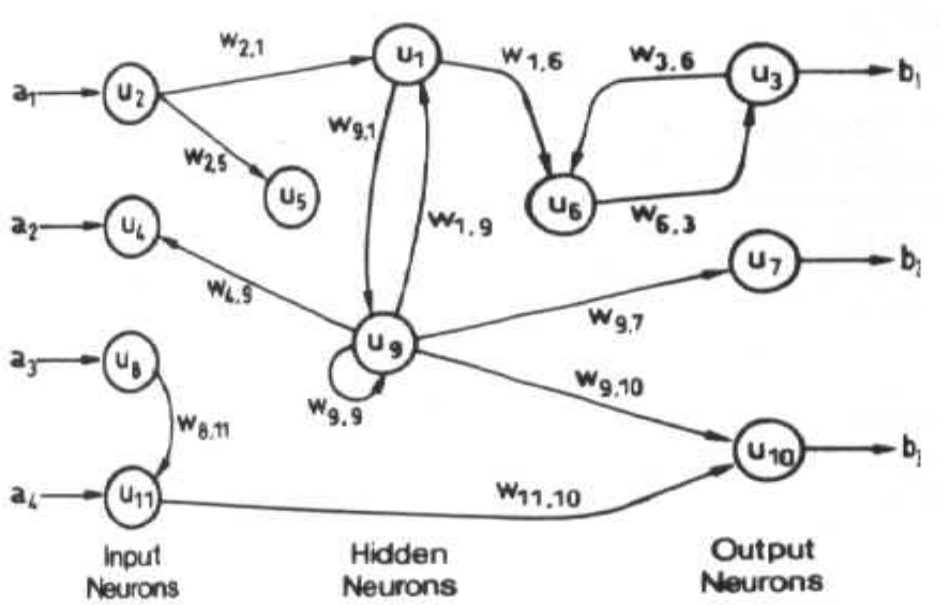
آکسون. ساده‌ترین این شبکه‌ها، شبکه‌های پرسپترون[1] هستند که در ادامه بیشتر در مورد آن بحث می‌گردد. شکل ۱-۵ نمونه‌ای از یک شبکه پیشخور را نشان می‌دهد.




*شکل ۱-۵: نمونه‌ای از یک شبکه پیش‌خور*

**شبکه‌های پسخور**[2]. تفاوت شبکه‌های برگشتی با شبکه‌های پیش‌خور در آن است که در شبکه‌های برگشتی حداقل یک سیگنال برگشتی از یک نرون به همان نرون یا نرون‌های همان لایه یا نرون‌های لایه‌های قبل وجود دارد و اگر نرونی دارای فیدبک باشد بدین مفهوم است که خروجی نرون در حال حاضر نه تنها به ورودی در آن لحظه بلکه به مقدار خروجی خود نرون، در لحظه‌ی گذشته نیز بستگی دارد. شبکه‌های برگشتی بهتر می‌توانند رفتار مربوط به ویژگی‌های زمانی و پویایی سیستم‌ها را نشان دهند. در این نوع شبکه‌ها که با توجه به ماهیت پویای مسئله طراحی می‌شوند، بعد از مرحله یادگیری شبکه نیز پارامترهای تغییر آورده و تصحیح می‌شوند. این شبکه‌ها پویا هستند؛ وضعیت آنها پیوسته در حال تغییر است تا اینکه آنها به یک نقطه تعادل برسند. آنها در این وضعیت تعادل باقی می‌مانند تا زمانی که ورودی تغییر کند و نیاز باشد تا تعادل تازه‌ای پیدا شود. ساده‌ترین این شبکه‌ها، شبکه **هاپفیلد** است. شکل شماره ۱-۶ نمونه‌ای از یک شبکه پسخور را نشان می‌دهد.

1. Perceptron
2. Feed-Back




*شکل ۱-۶: نمونه‌ای از یک شبکه برگشتی*

## مراحل طراحی یک شبکه عصبی مصنوعی

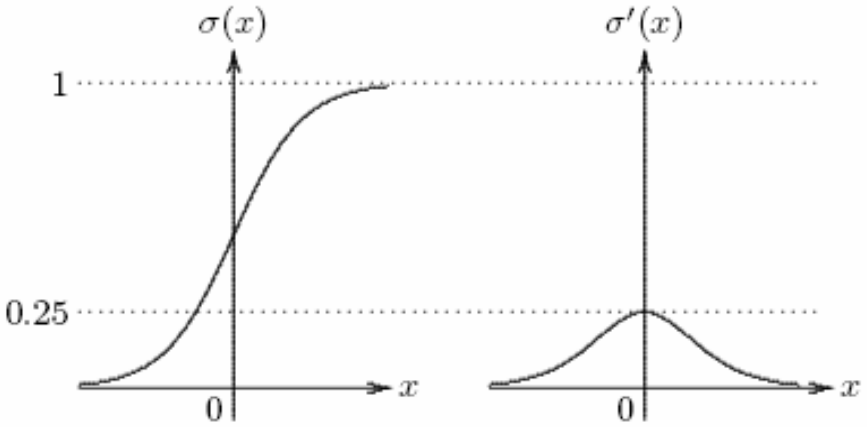
### مرحله ۱. طراحی معماری شبکه

این مرحله شامل تعیین تعداد لایه‌های موجود در شبکه، تعداد نرون‌های هر لایه، تعیین برگشت‌پذیر بودن یا نبودن شبکه و... است که با توجه به نوع مسأله تعیین می‌گردد. (برای مثال شبکه‌های برگشتی در اغلب موارد برای مسائل پویا کاربرد دارند و یا اینکه شبکه‌های پرسپترون پیش‌خور، برای نگاشت‌های غیرخطی کاربرد دارند.)

**نکته قابل توجه:** تعداد نرون‌های لایه ورودی از صورت مسأله‌ی مورد بررسی مشخص می‌گردد .به عبارت دیگر تحت انتخاب طراح مسأله نیست بلکه بستگی به روش حل مسأله موردنظر دارد. تعداد نرون‌های لایه خروجی بستگی به نوع جواب ما دارد. برای مثال چنانچه پاسخ ما به‌صورت یک عدد باشد یک نرون کافی است. تعداد لایه‌ها و تعداد نرون‌های لایه پنهان توسط کاربر تعیین می‌گردد اما در اکثر مسائل از یک تا سه لایه پنهان کفایت می‌کند. همچنین روش عملی‌ای برای تخمین تعداد نرون‌های لایه پنهان وجود ندارد به همین دلیل از روش‌های سعی و خطا (در حین آموزش) استفاده می‌شود تا به مقدار میانگین خطای مطلوب رسید.

## مرحله ۲. تعیین نوع تابع تبدیل

می‌توان برای اینکه خروجی خاصی تولید شود از یک تابع تبدیل استفاده کرد. این تابع بازه وسیعی از مقادیر ورودی را به مقدار خاصی نگاشت می‌کند. به‌عنوان مثال می‌توان هر مقدار خروجی را به مقدار باینری 0 و1 نگاشت کرد. انواع مختلفی از این توابع در ANN ها مورد استفاده قرار می‌گیرد ولی پرکاربدترین آنها، **تابع تبدیل سیگموید**[1] (مانند S) است. که به‌صورت زیر تعریف می‌شود:




ادعا شده است که فرکانس آتش نرون طبیعی به‌صورت تابعی شبیه به این تابع است. اما از دلایل عمده استفاده از این تابع این است که: تقریباً خطی، افزایشی و مشتق‌پذیر است، و در فرم بسته قابل نمایش است، مشتق‌گیری از آن ساده است و بازه ورودی ($+\infty$ ، $-\infty$) را به خروجی (۱ ، ۰ ) فشرده‌سازی می‌کند.

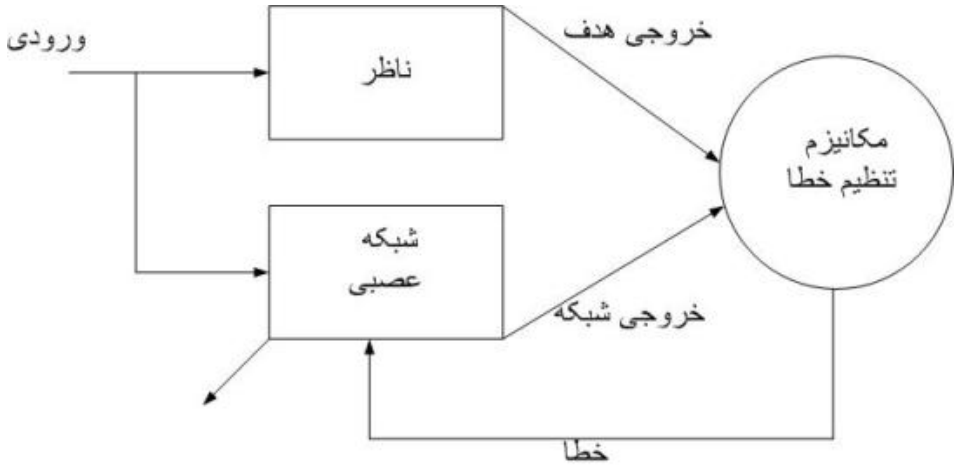
## مرحله ۳. آموزش شبکه

الگوریتم‌های یادگیری، روندهایی هستد که توسط آنها وزن‌های شبکه تنظیم می‌گردد. هدف از آموزش شبکه این است که شبکه قانون کار را یاد بگیرد و پس از آموزش به ازای هر ورودی، خروجی مناسب را ارائه دهد.

تاکنون بیش از ۱۰۰ نوع الگوریتم یادگیری به‌وجود آمده است که می‌توان آنها را به‌طورکلی به دو دسته وسیع تقسیم‌بندی کرد:

1. Sigmoid

**یادگیری نظارت شده یا با ناظر**[1] . در این نوع آموزش، به الگوریتم یادگیری مجموعه‌ای از زوج داده که به داده‌های یادگیری موسوم هستند، داده می‌شود. هر داده یادگیری شامل ورودی به شبکه و خروجی هدف است. پس از اعمال ورودی به شبکه، خروجی شبکه با خروجی هدف مقایسه می‌گردد و سپس خطای یادگیری محاسبه شده و از آن جهت تنظیم پارامترهای شبکه (وزن‌ها)، استفاده می‌گردد به‌گونه‌ای که اگر دفعه بعد به شبکه همان ورودی را دادیم، خروجی شبکه به خروجی هدف نزدیک گردد. شکل ۱-۷ یادگیری با ناظر را نشان می‌دهد.




*شکل ۱-۷*

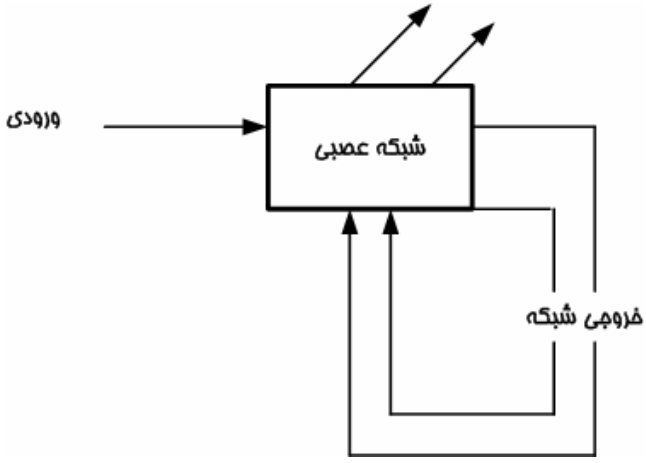
یادگیری تشدیدی[2] حالت خاصی از یادگیری با ناظر است که در آن به‌جای فراهم نمودن خروجی هدف به شبکه عددی که نشان‌دهنده میزان عملکرد شبکه است، ارائه می‌گردد. در یادگیری با ناظر، مقادیر خروجی هدف برای هر ورودی مفروض، کاملاً معلوم است ولی در بعضی مواقع اطلاعات کمی موجود می‌باشد. مثلاً به شبکه می‌توان گفت که خروجی‌اش ۵۰٪ درست است و... . در یادگیری با ناظر می‌گوییم جواب مطلوب برای x برابر t است. ولی در این الگوریتم‌های یادگیری می‌گوییم که شبکه چه قدر خوب به ورودی x جواب داده است.

1. Supervised learning
2. Reinforcement Learning

## یادگیری نظارت نشده یا بدون ناظر[1]

در این نوع یادگیری هیچ سیگنالی که اطلاعات را در مورد مطلوبیت شبکه به خود شبکه وارد نماید، وجود ندارد. به عبارت دیگر به شبکه گفته نمی‌شود که خروجی هدف چه است و یا اینکه جواب شبکه چقدر مطلوب است. در این حالت شبکه با دریافت اطلاعات ورودی، باید طبقه‌بندی‌ای بین الگوهای ورودی، شاخص‌های موجود در ورودی‌ها و ارتباط موجود بین الگوهای ورودی را پیدا کرده و در خروجی کد کند.

شکل زیر یادگیری بدون ناظر را نشان می‌دهد. یک مثال بسیار متداول از این نوع یادگیری، شبکه‌های خوشه‌بندی الگوهای ورودی است، بدون اینکه بدانیم کدام الگو به کدام خوشه تعلق دارد. خوشه‌ها در نهایت از روی تشابهات و عدم تشابهات بین الگوها ایجاد می‌گردند.

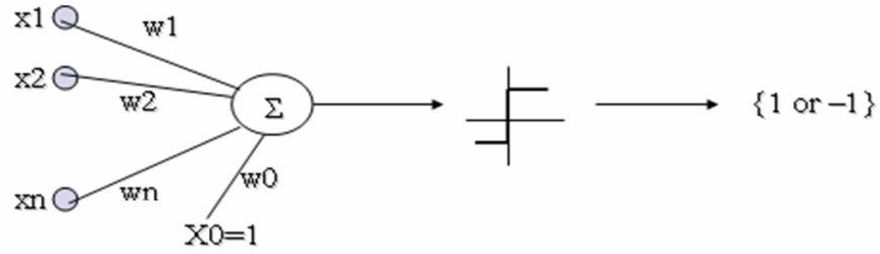



*یادگیری بدون ناظر*

1. Unsupervised Learning

## شبکه پرسپترون

نوعی از شبکه عصبی برمبنای یک واحد محاسباتی به نام پرسپترون ساخته می‌شود. یک پرسپترون برداری از ورودی‌های با مقادیر حقیقی را گرفته و یک ترکیب خطی از این ورودی‌ها را محاسبه می‌کند. اگر حاصل از یک مقدار آستانه بیشتر بود خروجی پرسپترون برابر با 1 و در این‌صورت معادل ۱- خواهد بود.



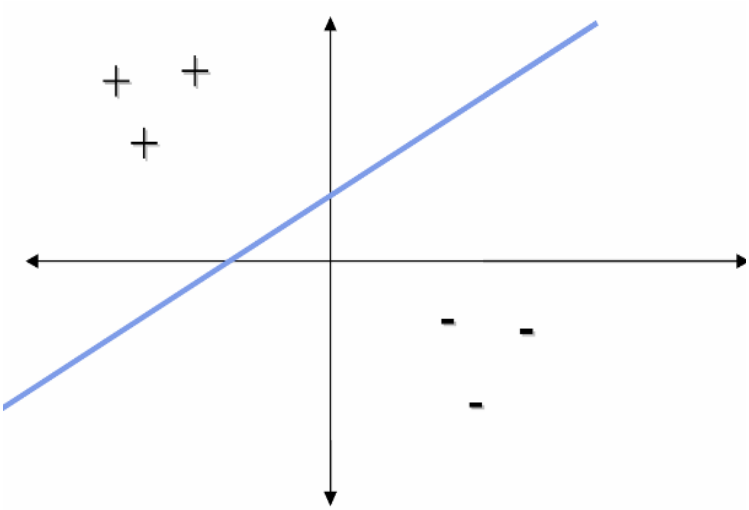

## یادگیری یک پرسپترون

خروجی پرسپترون توسط رابطه زیر مشخص می‌شود:

$$O(X_1, X_2, \ldots, X_n) = \begin{cases} 1 & \text{if } W_0X_0 + W_1X_1 + \ldots + W_nX_n > 0 \\ -1 & \text{otherwise} \end{cases}$$

**یادگیری پرسپترون عبارت است از:**
**پیدا کردن مقادیردرستی برای W**
**بنابراین فضای فرضیه H در یادگیری پرسپترون عبارت است ازمجموعه تمام مقادیر حقیقی ممکن برای بردارهای وزن**

پرسپترون را می‌توان به‌صورت یک سطح تصمیم Hyperplane در فضای n بعدی نمونه‌ها در نظر گرفت. پرسپترون برای نمونه‌های یک طرف صفحه مقدار 1 و برای مقادیر طرف دیگر مقدار ۱- به‌وجود می‌آورد.



## توابعی که پرسپترون قادر به یادگیری آنها می‌باشد

یک پرسپترون فقط قادر است مثال‌هایی را یاد بگیرد که به‌صورت خطی جدایی‌پذیر باشند. این‌گونه مثال‌ها مواردی هستند که به‌طور کامل توسط یک Hyperplane قابل جداسازی می‌باشند.

یک پرسپترون می‌تواند بسیاری از توابع بولی را نمایش دهد نظیر NOR ، OR ، AND ، NAND. اما نمی‌تواند XOR را نمایش دهد. در واقع هر تابع بولی را می‌توان با شبکه‌ای دوسطحی از پرسپترون‌ها نشان داد.

## الگوریتم‌های یادگیری پرسپترون

۱. مقادیری تصادفی به وزن‌ها نسبت می‌دهیم.

۲. پرسپترون را به تک‌تک مثال‌های آموزشی اعمال می‌کنیم. اگر مثال غلط ارزیابی شود مقادیر وزن‌های پرسپترون را تصحیح می‌کنیم.

۳. آیا تمامی مثال‌های آموزشی درست ارزیابی می‌شوند؟

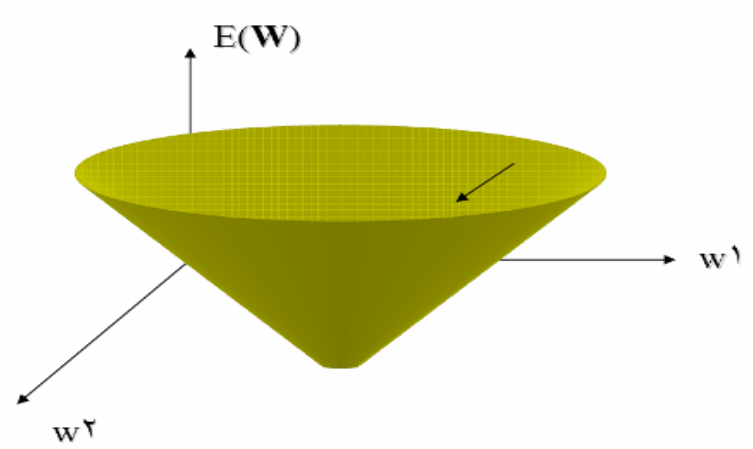
- بله. پایان الگوریتم
- خیر. به مرحله ۲ برمی‌گردیم

۴. وقتی که مثال‌ها به‌صورت خطی جدایی‌پذیر نباشند قانون پرسپترون همگرا نخواهد شد و برای غلبه بر این مشکل قانون دلتا استفاده می‌شود.

۵. ایده اصلی این قانون استفاده از gradient descent برای جستجو در فضای فرضیه وزن‌های ممکن می‌باشد. این قانون پایه روش Back Propagation است که برای آموزش شبکه با چندین نرون به هم متصل به‌کار می‌رود.

۶. همچنین این روش پایه‌ای برای انواع الگوریتم‌های یادگیری است که باید فضای فرضیه‌ای شامل فرضیه‌های مختلف پیوسته را جستجو کنند.

با توجه به نحوه تعریف E سطح خطا به‌صورت یک سهمی خواهد بود. ما به دنبال وزن‌هایی هستیم که حداقل خطا را داشته باشند. الگوریتم Gradient descent در فضای وزن‌ها به دنبال‌برداری می‌گردد که خطا را حداقل کند. این الگوریتم از یک مقدار دلخواه برای بردار وزن شروع کرده و در هر مرحله وزن‌ها را طوری تغییر می‌دهد که در جهت شیب کاهشی منحنی فوق خطا کاهش داده شود.




## مشکلات روش Gradient descent

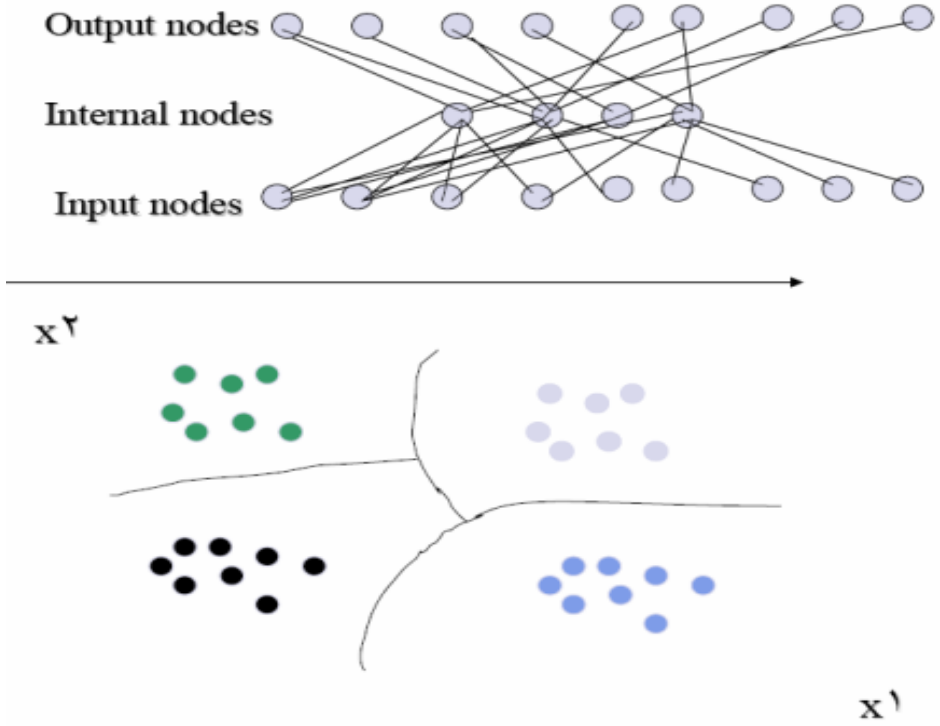
۱. ممکن است همگرا شدن به یک مقدار مینیمم زمان زیادی لازم داشته باشد.

۲. اگر در سطح خطا چندین مینیمم محلی وجود داشته باشد تضمینی وجود ندارد که الگوریتم مینیمم مطلق را پیدا بکند.

در ضمن این روش وقتی قابل استفاده است که :

- فضای فرضیه دارای فرضیه‌های پارامتریک پیوسته باشد.
- رابطه خطا قابل مشتق‌گیری باشد.

بر خلاف پرسپترون‌ها شبکه‌های چند لایه می‌توانند برای یادگیری مسائل غیرخطی و همچنین مسائلی با تصمیم‌گیری‌های متعدد به‌کار روند.




**برای اینکه بتوانیم فضای تصمیم گیری را به‌صورت غیرخطی از هم جدا بکنیم، لازم است تا هر سلول واحد را به‌صورت یک تابع غیرخطی تعریف نمائیم. مثالی از چنین سلولی می‌تواند یک واحد سیگموئید باشد.**

## الگوریتم Back propagation

۱. شبکه‌ای با $n_{in}$ گره ورودی، $n_{hidden}$ گره مخفی و $n_{out}$ گره خروجی ایجاد کنید.

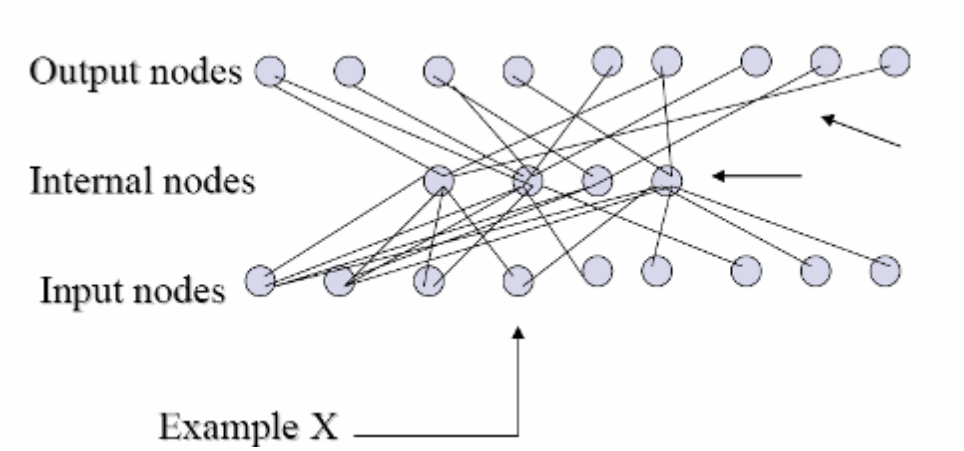
۲. همه وزن‌ها را با یک مقدار تصادفی کوچک عدد دهی کنید.

۳. تا رسیدن به شرط پایانی (کوچک شدن خطا) مراحل زیر را انجام دهید:

برای هر x متعلق به مثال‌های آموزشی:

- مثال x را به سمت جلو در شبکه انتشار دهید.

- خطای E را به سمت عقب در شبکه انتشار دهید.




۴. برای هر واحد خروجی جمله را به‌صورت زیر محاسبه کنید:

$$\delta_k = O_k(1 - O_k)(t_k - O_k)$$

۵. برای هر واحد مخفی جمله خطا را به‌صورت زیر محاسبه کنید:

$$\delta_h = O_h(1 - O_h)\sum_k^h W_{hk}\delta_k$$

۶. مقدار هر وزن را به‌صورت زیر تغییر دهید:

$$W_{ji} = W_{ji} + \Delta W_{ji}$$

که در آن :

$$\Delta W_{ji} = \vartheta \delta_j X_{ji}$$

معمولاً الگوریتم BP پیش از خاتمه هزاران بار با استفاده از همان داده‌های آموزشی تکرار می‌گردد شروط مختلفی را می‌توان برای خاتمه الگوریتم به کاربرد:

- توقف بعد از تکرار به دفعات معین
- توقف وقتی که خطا از یک مقدار تعیین شده کمتر شود.
- توقف وقتی که خطا در مثال‌های مجموعه تأیید از قاعده خاصی پیروی نماید.

اگر دفعات تکرار کم باشد خطا خواهیم داشت و اگر زیاد باشد مسئله over fitting رخ خواهد داد.

**چند نکته مهم درباره Back propagation:**

- این الگوریتم یک جستجوی gradient descent در فضای وزن‌ها انجام می‌دهد.
- ممکن است در یک مینیمم محلی گیر بیافتد.
- در عمل بسیار مؤثر بوده است.

برای پرهیز از مینیمم محلی روش‌های مختلفی وجود دارد:

- استفاده از stochastic gradient descent
- استفاده از شبکه‌های مختلف با مقادیر متفاوتی برای وزن‌های اولیه

## انواع مدل‌های پرسپترون چند لایه

پرسپترون چند لایه با یک لایه پنهان و تابع خروجی غیرخطی از نوع سیگموئید می‌باشد، که پرکاربردترین مدل از انواع پرسپترون‌های چند لایه است، ولیکن مدل‌های دیگری نیز وجود دارند.

این مدل‌ها براساس تغییراتی در مدل پرسپترون چند لایه استاندارد به منظور تسریع در آموزش شبکه به وجود آمده‌اند.

در روش تک لایه تغییر در اوزان توسط این معادله محاسبه می‌شود:

$$\eta x \delta = \Delta w_i$$

اصلاحی که در این معادله صورت گرفته است عبارت است از افزودن یک عبارت اضافی که از حاصل ضرب مقدار ثابت $\alpha$ در مقدار کنونی $\Delta w$ به‌دست می‌آید. یعنی:

$$(\Delta w)_{k+1} = \eta x \delta + \alpha (\Delta w)_k$$

انتخاب مقدار مناسبی برای $\alpha$ به روش سعی و خطا صورت می‌گیرد. البته معمول آن است که مقداری برای $\alpha$ انتخاب شود که از $\eta x \delta$ کوچکتر باشد.

تغییر دیگر در مدل استاندارد این است که به جای محدوده ۰ تا ۱ برای خروجی از فاصله ۱- تا۱+ برای آن استفاده می‌شود. انتظار می‌رود که انجام این کار باعث تسریع آموزش شود

زیرا در این حالت تابع نامتقارن است یعنی f(-x)=-f(x) می‌باشد در عوض در این حالت خروجی‌های تولید شده دارای میانگین صفر می‌باشد در عوض در این حالت خروجی‌های تولید شده دارای میانگین صفر می‌باشند حال آن که در تابع سیگموئید استاندارد، تمامی مقادیر خروجی‌ها مثبت هستند. برای رفع این مشکل تابع سیگموئید، از تابع سیگموئید اصلاح شده استفاده می‌شود:

$$y = \frac{2}{1+e^{-x}} - 1 = \frac{(1-e^{-x})}{(1+e^{-x})}$$

درصورتی‌که از این تابع که تانژانت هیپربولیک نامیده می‌شود، استفاده شود، قاعده پس انتشار[1] باید به‌گونه‌ای اصلاح شود که مشتق y نسبت به x به شکل زیر در می‌آید:

$$Y' = 2y(1-y)$$

یعنی می‌توان ضریب یادگیری $\eta$ را دو برابر کرد.

مدل‌های دیگری که به‌عنوان تغییر شکلی از پس انتشار به کار می‌روند، مدل‌هایی هستند که در آنها بجای تابع سیگموئید از تابع‌های مثلثاتی نظیر $\sin(x)$ یا $\tan(x)$ استفاده می‌شود. وقتی ورودی ها به‌صورت سیگنال‌هایی باشند که توسط سری‌های فوریه نشان داده شوند، استفاده از این مدل‌ها بر مدل اصلی ارجح است.

## شبکه‌های تابع شعاع مدار[2] (RBF)

یکی از مهم‌ترین گونه‌های شبکه عصبی تابع شعاع مدار (RBF) می‌باشد. این شبکه با توجه به کاربردهای متنوعش به یکی از معروفترین شبکه‌های عصبی تبدیل شده است و مهمترین رقیب پرسپترون چند لایه محسوب می‌شود. این شبکه‌ها بیشترین الهام را از تکنیک‌های آماری طبقه‌بندی الگوها گرفته‌اند.

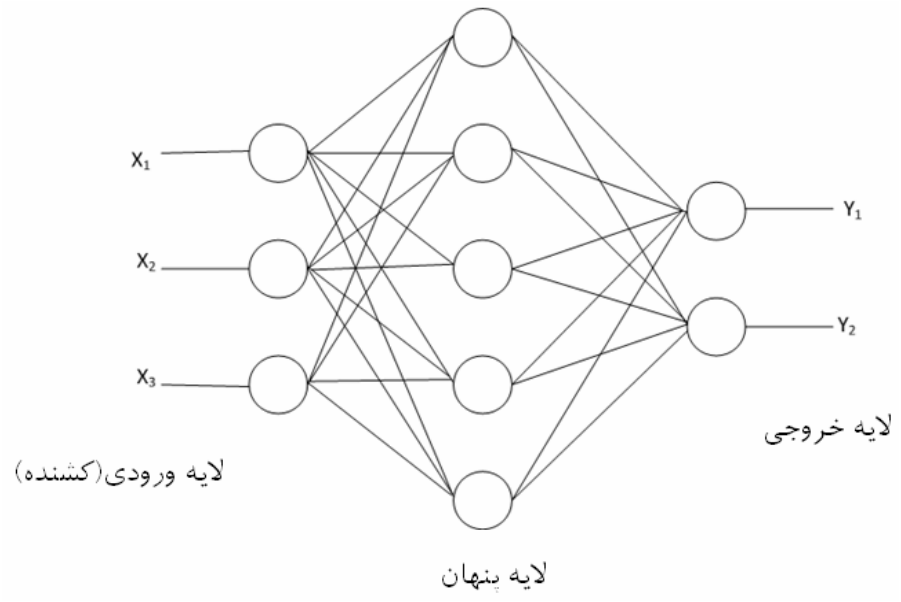
1. Back Propagation
2. Radial Basis function

## معماری شبکه

معماری اصلی RBF متشکل از یک شبکه سه لایه می‌باشد که در شکل صفحهٔ بعد ملاحظه می‌کنید. لایه ورودی فقط یک لایه کشنده است و در آن هیچ پردازشی صورت نمی‌گیرد. لایه دوم یا لایه پنهان، یک انطباق غیرخطی ما بین فضای ورودی و یک فضا (معمولاً) با بعد بزرگتر برقرار می‌کند که در آن الگوها به صورت تفکیک‌پذیر خطی درمی‌آیند. سرانجام لایه سوم، جمع وزنی را به همراه یک خروجی خطی تولید می‌کند. درصورتی‌که از RBF برای تقریب تابع استفاده شود، چنین خروجی‌ای مفید خواهد بود ولی درصورتی‌که نیاز باشد طبقه‌بندی الگوها انجام شود، آنگاه یک محدودکننده سخت یا یک تابع سیگموئید را می‌توان برروی عصب‌های خروجی قرار داد تا مقادیر خروجی ۰ یا ۱ تولید شوند.

خصوصیت منحصربه‌فرد RBF پردازشی است که در لایه پنهان انجام می‌گیرد. ایده اصلی آن است که الگوهای فضای ورودی، تشکیل خوشه دهند. درصورتی‌که مراکز این خوشه‌ها مشخص باشد، می‌توان فاصله از مرکز خوشه را اندازه گرفت. به‌علاوه این اندازه‌گیری فاصله به‌صورت غیرخطی انجام می‌گیرد، لذا درصورتی‌که الگویی در ناحیه مجاور مرکز یک خوشه قرار داشته باشد مقداری نزدیک به ۱ تولید می‌شود. در خارج از این ناحیه، مقدار به‌دست آمده به‌طور قابل ملاحظه‌ای کاهش می‌یابد. نکته مهم آن است که که این ناحیه به‌صورت شعاعی در اطراف مرکز خوشه متقارن است، بنابراین تابع غیرخطی به‌صورت تابع شناخته شده شعاع مدار در می‌آید. معمولی‌ترین شکل تابع شعاع مدار بدین‌صورت است:

$$\emptyset(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right)$$




در یک RBF ، r برابر مقدار عددی فاصله از مرکز خوشه می‌باشد. معمولاً فاصله اندازه‌گیری شده تا مرکز خوشه، از نوع فاصله اقلیدسی است. برای هر عصب موجود در لایه پنهان، وزن‌ها، مختصات مرکز خوشه را نشان می‌دهند. بنابراین زمانی که عصب یک الگوی ورودی x را دریافت می‌کند، فاصله مزبور با استفاده از معادله زیر به‌دست می‌آید:

$$r_j = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

بنابراین خروجی عصب j در لایه پنهان به این شکل است:

$$\emptyset_j = \exp\left( -\frac{\sum_{n=1}^{n} (x_i - w_{ij})^2}{2\sigma^2} \right)$$

متغیر $\sigma$ به‌عنوان عرض یا شعاع منحنی نرمال تعریف می‌شود و گاهی اوقات الزاماً به‌صورت تجربی تعیین می‌شود. زمانی که فاصله از مرکز منحنی نرمال به$\sigma$ می‌رسد، خروجی از ۱ به ۰,۶ تنزل می‌یابد.

مثال مشهوری که اغلب به‌منظور نشان دادن نحوه کارکردن شبکه RBF برروی یک تابع تفکیک‌پذیر غیرخطی ارائه می‌شود، مسأله OR انحصاری است.

# آموزش شبکه RBF

## لایه پنهان

لایه پنهان یک شبکه RBF دارای واحدهایی است که وزن‌دار بوده و اوزان آن متناظر با بردار نشان‌دهنده‌ی مرکز خوشه می‌باشد. وزن‌ها را می‌توان با استفاده از شیوه‌های سنتی نظیر الگوریتم k میانگین[1] یا روش‌های مبتنی بر الگوریتم کوهونن به‌دست آورد. در اینجا آموزش به‌صورت غیرنظارتی انجام می‌شود ولی تعداد خوشه‌های موردنظر (k) پیشاپیش انتخاب می‌شوند و این الگوریتم‌ها بهترین برازش را برای خوشه‌ها به‌دست می‌آورند. در اینجا الگوریتم k میانگین را به‌صورت مختصر شرح می‌دهیم:

نخست به‌صورت تصادفی k نقطه در فضای الگوها قرار داده می‌شود، سپس برای هر داده از مجموعه‌ی آموزش، فاصله از تمامی K مرکز محاسبه شده و نزدیکترین مرکز برای هر داده انتخاب می‌شود. بدین‌ترتیب یک طبقه‌بندی ابتدایی به دست آمده که در آن هر داده به یکی از کلاس‌های ۱ تا k تخصیص داده می‌شود. سپس برای تمامی داده‌های تخصیص داده شده به کلاس ۱، مقادیر میانگین هر یک از مختصه‌ها محاسبه می‌شود. این مقادیر، مختصه‌های مرکز جدید مربوط به کلاس ۱ خواهند بود. این رویه برای تمامی کلاس‌های دیگر نیز تکرار می‌شود. اکنون K مرکز جدید داریم. کلیه مراحل قبلی برای داده‌ها مجدداً تکرار می‌شود تا جایی که دیگر تغییری مشاهده نشود. مجموع فاصله‌ها محاسبه شده و رویه زمانی متوقف می‌شود. پس از به‌دست آوردن مراکز خوشه‌ها توسط یکی از الگوریتم‌های یاد شده، مرحله بعد تعیین شعاع منحنی نرمال (گوسی) می‌باشد. معمولاً این شعاع توسط الگوریتم نزدیکترین P همسایه[2] به‌دست می‌آید. عدد P انتخاب شده و برای هر مرکز p مرکز نزدیکتر مشخص می‌شوند. جذر متوسط مجذور فاصله بین مرکز خوشه کنونی و P همسایه این مرکز محاسبه می شود و این مقدار به‌عنوان $\sigma_j$ انتخاب می‌گردد. درصورتی‌که مرکز خوشه کنونی $c_j$ باشد، مقدار $\sigma_j$ به شکل زیر محاسبه می‌شود:

$$\sigma_j = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(c_k - c_i)^2}$$

1. K-mean
2. Nearest neighbor algoritm

معمولاً برای p ، مقدار ۲ انتخاب می‌شود لذا در این حالت $\sigma$ برابر میانگین فاصله از مرکز دو خوشه (نزدیکترین) خواهد بود.

با استفاده از این شیوه برای آموزش لایه پنهان، تابع OR انحصاری را می‌توان با به‌کارگیری حداقل چهار واحد پنهان اجرا نمود. درصورتی‌که برای اجرای این تابع از بیش از ۴ واحد استفاده می‌شود، واحدهای اضافی، تعداد مراکز را دو برابر می‌نمایند لذا در قابلیت تمیز شبکه نقشی نخواهند داشت. با فرض اینکه در لایه پنهان، چهار واحد وجود داشته باشد، هر واحد در مرکز یکی از چهار الگوی ورودی با نام‌های ۰۰، ۰۱، ۱۰، ۱۱ قرار داده می‌شود. از الگوریتم نزدیکترین p همسایه با اتخاذ p=2 برای به‌دست آوردن اندازه شعاع‌ها استفاده می‌شود. در هر یک از عصب‌ها، فاصله از هر یک از ۳ عصب دیگر، برابر ۱ ، ۱ و ۱/۴۱۴ می‌باشد. بنابراین دو مرکز خوشه نزدیکتر در فاصله ۱ قرار دارند. با به‌کارگیری فرمول میانگین مجذور فاصله، برای هر عصب شعاع ۱ به‌دست می‌آید.

با به‌کارگیری مقادیر به‌دست آمده برای مراکز و شعاع‌ها، خروجی لایه پنهان برای هر یک از چهار الگوی ورودی به‌صورت زیر خواهد بود:

| ورودی | عصب ۱ | عصب ۲ | عصب ۳ | عصب ۴ |
|---|---|---|---|---|
| ۰۰ | ۰/۶ | ۰/۴ | ۱/۰ | ۰/۶ |
| ۰۱ | ۰/۴ | ۰/۶ | ۰/۶ | ۱/۰ |
| ۱۰ | ۱/۰ | ۰/۶ | ۰/۶ | ۰/۴ |
| ۱۱ | ۰/۶ | ۱/۰ | ۰/۴ | ۰/۶ |

## لایه خروجی

پس از آموزش لایه پنهان توسط الگوریتم‌های یادگیری غیرنظارتی، مرحله نهایی آموزش لایه خروجی با استفاده از یک تکنیک استاندارد کاهش شیب، نظیر الگوریتم LMS که آدالاین به‌کار رفت، انجام می‌گیرد. در مثال تابع OR انحصاری که در بالا عنوان شد، مجموعه‌ای مناسب برای وزن‌ها می‌تواند به‌صورت ۱+ ، ۱− ، ۱− و ۱+ باشد. با به‌کارگیری این وزن‌ها، مقدار net و خروجی به شکل زیر خواهد بود:

| خروجی | net | عصب ۴ | عصب ۳ | عصب ۲ | عصب ۱ | ورودی |
|---|---|---|---|---|---|---|
| ۰ | -۰/۲ | ۰/۶ | ۱/۰ | ۰/۴ | ۰/۶ | ۰۰ |
| ۱ | ۰/۲ | ۱/۰ | ۰/۶ | ۰/۶ | ۰/۴ | ۰۱ |
| ۱ | ۰/۲ | ۰/۴ | ۰/۶ | ۰/۶ | ۱/۰ | ۱۰ |
| ۰ | -۰/۲ | ۰/۶ | ۰/۴ | ۱/۰ | ۰/۶ | ۱۱ |

## مزایای یک RBF

RBF مزایای زیادی در مقابل پرسپترون چند لایه (MLP) دارد. یکی از مزیت‌های آن سرعت بیشتر و ایجاد محدوده‌های تصمیم‌گیری بهتر است. مزیت دیگر RBF این است که در این شبکه تعبیر و تفسیر لایه خروجی به مراتب آسان‌تر از یک MLP انجام می‌شود.

## برخی کاربردهای شبکه‌های عصبی مصنوعی

کاربردهای کلی شبکه عصبی را می‌توان به سه دسته تقسیم کرد:

- دسته‌بندی و شناسایی الگو
- پیش‌بینی
- مدل‌سازی

**دسته‌بندی و شناسایی الگو**: می‌توان با استفاده از این شبکه‌ها انواع الگوها را دسته‌بندی و از هم تفکیک کرد.

- دسته‌بندی نقاط خارج از کنترل در کنترل کفیت
- تفکیک و دسته‌بندی نظرات خبرگان از عامه در سیستم پشتیبان تصمیم‌گیری (DSS)
- دسته‌بندی بهینه ماشین‌آلات

**پیش‌بینی**: این‌گونه شبکه‌ها به‌گونه‌ای آموزش دیده‌اند که براساس یادگیری و حفظ تجارب، قادر به پیش‌بینی آینده هستند.

- شبکه‌های پیش‌بینی قیمت نفت و بازار بورس
- شبکه‌های پیش‌بینی‌کننده در مباحث کنترل موجودی، کنترل کیفیت و برنامه‌ریزی تعمیرات

**مدلسازی**: این‌گونه شبکه‌ها به‌طور گسترده‌ای در مسائل برنامه‌ریزی تولید و OR کاربرد دارند. شبکه‌هایی طریحی شده‌اند که قادر به جستجوی نقطه بهینه سراسری در زمینه برنامه‌ریزی TPS ، Job-shop Schedul و... هستند.

## بهینه سازی:

- در سیستم‌های کنترلی
- در سیستم‌های مدیریت، تخصیص و تسهیم منابع
- در سیستم‌های مالی، می‌توان شبکه‌های عصبی (علی‌الخصوص شبکه‌های برگشتگی) استفاده کرد.

# فصل دوم

# راهنمای کاربر

# بخش اول

## شبکه‌های عصبی[1] در SPSS

شبکه‌های عصبی به علت قدرت، انعطاف‌پذیری و سهولت استفاده شان، ابزار برتر در بسیاری از کاربردهای مربوط به فرآیندهای پیش‌بینی از طریق تحلیل داده‌ها هستند. شبکه‌های عصبی پیشگو خصوصاً در مواردی که فرآیند پیچیده است مورد استفاده قرار می‌گیرند، مانند:

- پیشگویی تقاضای مشتری برای هزینه حمل و تولید مؤثر
- تشخیص خانوار مناسب از یک mailing list جهت فرستادن پیشنهاد از طریق پیش‌بینی احتمال پاسخ مثبت به بازاریابی مستقیم پستی
- امتیازدهی به متقاضی جهت تعیین ریسک حساب‌های اعتباری
- تشخیص تراکنش‌های گمراه‌کننده در پایگاه داده تقاضاهای بیمه

شبکه عصبی مانند پرسپترون چند لایه[2] (MLP) و تابع شعاع مدار[3] (RBF) که در پیش‌بینی استفاده می‌شوند، به‌گونه‌ای عمل می‌کنند که نتایج پیش‌بینی حاصل از مدل می‌تواند با مقادیر متغیر هدف مقایسه شود. شبکه عصبی این امکان را به شما می‌دهد که شبکه‌های MLP و RBF را تنظیم کرده و نتایج مدل را برای امتیازدهی ذخیره نمایید.

## ساختار شبکه عصبی

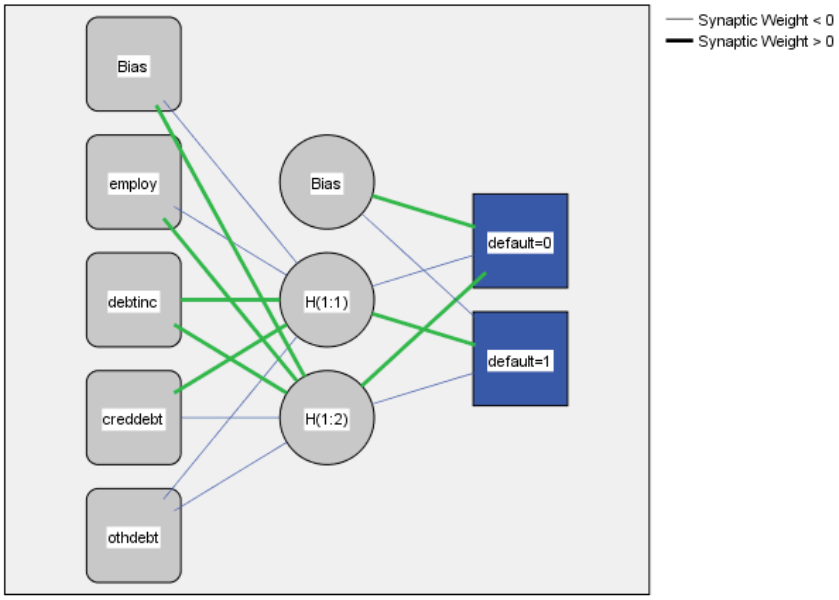
با وجود آنکه شبکه‌های عصبی به پیش‌فرض‌ها و ساختار مدلی نیاز کمی دارد اما درک مفهوم کلی معماری شبکه مفید است. شبکه پرسپترون چند لایه یا تابع شعاع مدار، تابعی از پیش‌بینی‌کننده‌ها (که همچنین از آنها به‌عنوان متغیرهای مستقل یا ورودی‌ها یاد می‌شود) است که خطای پیش‌بینی متغیرهای هدف (خروجی‌ها) را حداقل می‌کند.

به مجموعه داده‌های موجود در فایل bankloan.sav که همراه با نرم‌افزار است و در آن هدف شناسایی بدحسابان بالقوه در میان تعداد زیادی از متقاضیان وام است توجه نمایید.

---

1. Neural Network
2. Multilayer Perceptron
3. Redial Basis Function

شبکه MLP و RBF که برای این مسئله استفاده می‌شود تابعی از اندازه‌گیری‌هاست که خطای پیش‌بینی بدحسابی را حداقل می‌کند. شکل زیر بیانگر رابطه موجود میان ساختار این تابع است.




این ساختار معروف به معماری پیش‌خور[1] (feedforward architecture) است، چون رابطه‌های درون شبکه از لایه ورودی به لایه خروجی بدون هیچ بازگشتی، به جلو جریان دارد. در این شکل:

- لایه ورودی شامل پیش‌بینی‌کننده است
- لایه ورودی شامل گره‌ها یا واحدهای غیرقابل مشاهده است ارزش هر واحد پنهان تابعی از پیش بینی‌کننده است. ساختار دقیق تابع به دو عامل وابسته است: نوع شبکه و مشخصه‌های قابل کنترل توسط کاربر.
- لایه خروجی شامل عکس‌العمل‌ها می‌باشد.

1. feedforward architecture

از آنجایی که سابقه بدحسابی یک متغیر مطلق با دو دسته‌بندی است، با دو متغیر نشانگر ثبت می‌شود. هر واحد خروجی به نحوی تابعی از واحدهای پنهان هستند. مجدداً ساختار دقیق تابع به دو عامل نوع شبکه و مشخصه‌های قابل کنترل توسط کاربر وابسته است.

شبکه MLP امکان داشتن دو لایه پنهان را فراهم می‌کند در این صورت هر واحد در دومین لایه پنهان، تابعی از واحدهای موجود در لایه پنهان اول بوده و هر پاسخ، تابع واحدهای لایه پنهان دوم می‌باشد.

# بخش دوم

## پرسپترون چند لایه

روش پرسپترون چند لایه (MLP) یک مدل پیش‌بینی‌کننده برای یک یا چند متغیر وابسته (هدف) براساس مقادیر متغیرهای پیش‌بینی‌کننده، فراهم می‌کند.

**مثال.** در دو مثال زیر از روش MLP استفاده شده است.

یک مأمور بخش اعطاء وام که در بانک مشغول به فعالیت می‌باشد، نیازمند آن است تا بتواند ویژگی‌ها و خصوصیات افرادی که ممکن است در باز پرداخت وام غفلت و تأخیر ورزند را شناخته تا با استفاده از آنها میزان ریسک حساب‌های اعتباری آنان را بشناسد. به‌وسیله نمونه‌ای از مشتریان قبلی، وی می‌تواند یک پرسپترون چند لایه را آموزش داده، تحلیل‌ها را با استفاده از یک "نمونه جدا نگه داشته شده[1]" مشتریان سابق تصدیق کرده و سپس از شبکه برای طبقه‌بندی ریسک اعتبار مشتریان آینده استفاده نماید.

یک سیستم بیمارستانی تمایل دارد هزینه‌ها و مدت اقامت بیمارانی که برای درمان انفاکتوس میوکارد (MI حمله قلبی) پذیرش می‌شوند پیگیری کند. به‌دست آوردن تخمین دقیق این مقادیر به مدیر این امکان را می‌دهد که تخت‌های خالی موجود را در زمانی که بیماران درمان می‌شوند، مدیریت کند. با استفاده از نمونه‌ی پرونده‌های بایگانی بیماران MI که قبلاً درمان شده‌اند مدیر می‌تواند یک شبکه را برای پیش‌بینی هزینه و مدت اقامت آموزش دهد.

## متغیرهای وابسته

متغیرهای وابسته می‌توانند به این صورت باشند:

- **اسمی**[2]. در این نوع مقیاس برای اندازه‌گیری متغیر از اسامی، حروف و... استفاده می‌شود مانند اسم افراد، اسم شرکت‌ها و...

1. Hold out sample
2. Nominal

- **مقیاس ترتیبی**[1]. در این نوع مقیاس ترتیبی در میان سطوح متغیر وجود دارد مانند طیف پنج گزینه‌ای خیلی‌کم، کم، متوسط، زیاد و خیلی زیاد برای رضایت کارکنان یا طیف سه گزینه‌ای درجه ۱، ۲ و ۳ برای کیفیت محصول.
- **مقیاس فاصله‌ای**. صفر آن قراردادی بوده و فاصله بین دو واحد متوالی آن مقدار ثابتی است. مقیاس‌های سانتی‌گراد و فارنهایت نمونه‌هایی از این مقیاس می‌باشند.
- **مقیاس نسبی (نسبتی)**. مشابه مقیاس فاصله‌ای است با این تفاوت که نقطه صفر آن واقعی می‌باشد مانند مقیاس‌های سانتی‌متر و اینچ.

در این مراحل فرض بر آن است که سطح مناسب اندازه‌گیری برای تمام متغیرهای وابسته، نسبت داده شده است. به هر حال شما می‌توانید موقتاً سطوح اندازه‌گیری برای یک متغیر را با کلیک راست برروی متغیر در لیست آن و انتخاب یک سطح اندازه‌گیری از فهرست تغییر دهید.

یک نشانه‌گر درکنار هر متغیر در لیست متغیر، سطح اندازه‌گیری و نوع داده آن را مشخص می‌کند.

| Measurement Level | Data Type | | | |
|---|---|---|---|---|
| | Numeric | String | Date | Time |
| Scale |  | n/a |  |  |
| Ordinal |  |  |  |  |
| Nominal |  |  |  |  |

**متغیر پیش‌بینی‌کننده**[2]. پیش‌بینی‌کننده‌ها می‌توانند به‌صورت مطلق یا نسبی تعیین شوند.

**کدگذاری متغیر مطلق**[3]. فرایند به‌صورت موقتی، متغیرهای وابسته و پیش‌بینی‌کننده مطلق را با استفاده از یکی از کدهای c تا اتمام مراحل، کدگذاری مجدد می‌کند. اگر c دسته از یک متغیر وجود داشته باشد، به‌عنوان بردارهای c ذخیره می‌شود، اولین نوع بردار (1,0,…,0) بعدی (0,1,0,…,0) و بالاخره آخرین نوع (0,0,…,1) است.

این مدل کدگذاری تعداد وزن‌های سیناپسی را افزایش داده و باعث آموزش آهسته‌تر می‌شود. به هر حال هرچه روش‌های کدگذاری فشرده‌تر باشند معمولاً به انطباق ضعیف‌تری با

1. Ordinal
2. Predictor variable
3. Categorical variable coding

شبکه عصبی منجر خواهد شد. اگر شبکه‌ی آموزشی شما پیشرفت بسیار آهسته‌ای دارد سعی کنید تعداد پیش‌بینی‌کننده‌های مطلق خود را به‌وسیله ترکیب کردن با دسته‌های مشابه و یا حذف دسته‌های خیلی نادر، کاهش دهید.

حتی اگر یک نمونه آزمایش[1] یا "نمونه جدا نگه داشته شده" مشخص شده باشد، همه‌ی کدگذاری ۱ از c برپایه داده آموزش است. اگر نمونه جدا نگه داشته شده یا نمونه آزمایش[2] شامل مواردی باشد که دسته‌های پیش‌بینی‌کننده آن در داده آموزش ارائه نشده است، از آنها در برنامه یا امتیازدهی استفاده نخواهند شد. چنانچه نمونه‌های آزمایش یا جدا از هم نگه داشته شده شامل حالتی از دسته‌های متغیر وابسته باشد که در داده‌های آموزش آورده نشده است، این حالت‌ها در فرایند استفاده نشده اما ممکن است امتیازدهی شوند.

**مقیاس‌بندی مجدد**[3]. متغیرها و متغیرهای کمکی[4] وابسته به مقیاس جهت بهبود آموزش شبکه به‌صورت قراردادی مقیاس‌بندی مجدد می‌شوند. حتی در صورت وجود نمونه‌های آزمایش و جدا نگه داشته شده نیز مقیاس‌بندی‌های مجدد براساس داده‌های آموزش انجام می‌شود. بدین‌معنی که براساس نوع مقیاس‌بندی مجدد، میانگین، انحراف معیار، مقدار حداقل یا حداکثر یک متغیر کمکی یا متغیر وابسته تنها با استفاده از داده‌های آموزش محاسبه می‌شوند. تبعیت کردن این متغیرهای کمکی یا وابسته از یک توزیع مشابه در نمونه‌های آموزش، آزمایش و جدا نگه داشته شده حین تفکیک یک متغیر خاص مهم است.

**وزن‌های فراوانی**[5]. وزن‌های فراوانی در این روش نادیده گرفته می‌شود.

**قابلیت تکرارپذیری نتایج**[6]. اگر می‌خواهید دقیقاً نتایج را تکرار کنید، از همان مقدار اولیه برای تولید ارقام تصادفی و همان داده ومتغیر استفاده کنید. علاوه براین نیاز است از همان تنظیمات نیز استفاده نمایید. جزئیات بیشتر این مسئله به شرح زیر است:

- **تولید اعداد تصادفی**. در این برنامه برای تولید اعداد تصادفی از این موارد بهره جسته می‌شود: تخصیص تصادفی قسمت‌ها، پیش‌نمونه‌گیری تصادفی جهت مقداردهی اولیه

1. Training sample
2. Testing sample
3. Rescaling
4. covariate
5. Frequency weights
6. Replicating results

وزن‌های سیناپسی، پیش‌نمونه‌گیری تصادفی برای انتخاب ساختار و الگوریتم خودکار شبیه‌سازی شده برای استفاده در وزن‌دهی اولیه.

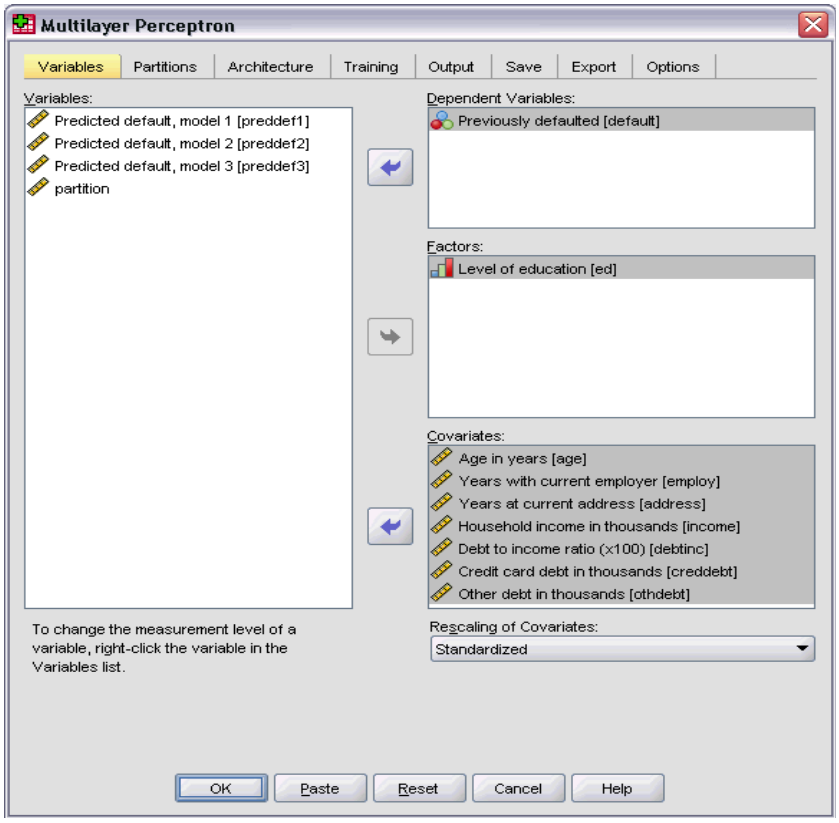
- به‌منظور باز تولید نتایج تصادفی به‌دست آمده می‌بایست، پیش از هر بار اجرای روش MLP از همان مقادیر اولیه برای تولید اعداد تصادفی استفاده نمود.
- **مرتبه حالت**[1] . روش‌های آموزش روی خط[2] و دسته‌ای کوچک[3] مستقیماً به مرتبه حالت وابسته‌اند، حتی آموزش دسته‌ای نیز از آنجایی که وزن‌دهی اولیه سیناپسی، به‌وسیله پیش‌نمونه‌گیری از کل اطلاعات در دسترس انجام می‌گیرد، به مرتبه حالت وابستگی پیدا می‌کند. به منظور حداقل نمودن اثرات مرتبه‌ای می‌باید به شکل تصادفی حالات مختلف را به هم مرتب نمود. جهت بررسی ثبات مراحل مطرح شده، نیاز به راه‌حل‌های متفاوت با طبقه‌بندی‌های مختلف تصادفی خواهید داشت. در شرایطی که اندازه فایل خیلی بزرگ باشد، چندین اجرا برروی یک نمونه از حالت‌هایی که در مراتب تصادفی ذخیره شده‌اند، کافی است.
- **مرتبه متغیر**[4] . نتایج حاصل از الگوهای متفاوت که تحت‌تأثیر مقادیر اولیه متغیرها قرار دادن، با تغییر در مرتبه متغیرها، تغییر می‌نمایند. در موارد مربوط به ضریب و متغیر کمکی نیز همان‌طورکه در مرتبه حالت وجود داشت، برای ارزیابی ثبات راه حل ارائه شده می‌توانید از مرتبه‌های متفاوتی برای متغیرها استفاده کنید. (با drag کردن و قرار دادن در لیست متغیرهای کمکی و ضرایب، این کار به راحتی انجام می‌شود.)

## ساخت یک شبکه پرسپترون چندلایه

از منو انتخاب کنید:

analyze→neural network→multilayer perceptron

1. Case order
2. online
3. Mini-batch
4. Variable order




*شکل ۲-۱ : پرسپترون چندلایه : نوار متغیر*

- حداقل یک متغیر وابسته انتخاب کنید
- حداقل یک ضریب یا متغیر کمکی انتخاب نمایید
- می‌توانید برروی نوار variable روش مقیاس‌بندی مجدد متغیرهای کمکی را تغییر دهید.

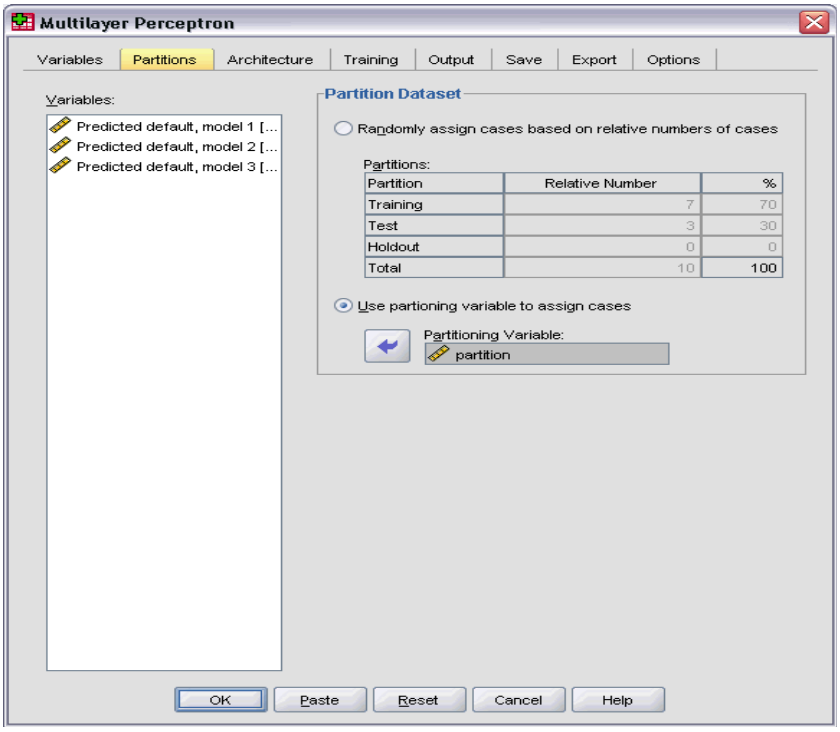
گزینه‌ها عبارتند از:

- استاندارد شده : میانگین را تفریق کرده و بر انحراف معیار تقسیم نمایید. $\frac{x - mean}{s}$
- نرمال شده[1] : مقدار مینیمم را تفریق کرده و بر دامنه تقسیم نمایید. $\frac{x - min}{max - min}$
- مقادیر نرمال شده بین ۰ و ۱ قرار می‌گیرند.

1. normalized

- نرمال تنظیم شده[1] : نمونه تنظیم شده‌ای است از مقادیر نرمال شده که مقدار آن بین ۱- و ۱ قرار می‌گیرد. $\frac{2(x-\min)}{\max-\min}-1$
- هیچکدام[2] : بدون مقیاس‌بندی مجدد متغیرهای کمکی

## Partitions (تفکیک کردن)




*شکل ۲-۲ : پرسپترون چندلایه : نوار Partitions*

**تفکیک مجموعه داده‌ها**[3]: در این قسمت به تشریح روش‌های تفکیک‌کننده مجموع داده‌های فعال به نمونه‌های آزمایش، جدا نگه داشته شده و نمونه‌های آموزش می‌پردازیم.

**"نمونه آزمایش"** شامل داده‌های ثبت شده‌ای است که در آموزش شبکه عصبی مورد استفاده قرار می‌گیرد. جهت ساخت مدل می‌بایست بخشی از موارد موجود در مجموعه داده‌ها را به

1. Adjusted Normalized
2. None
3. Partition dataset

نمونه آموزش، تخصیص دهیم. نمونه آزمایش یک مجموعه مستقل از داده‌های ذخیره شده است که از آن جهت پیدا کردن خطاهای رخ داده در حین آموزش استفاده می‌گردد و این امر از انجام آموزش بیش از حد جلوگیری می‌نماید.

توصیه اکید بر این است که یک نمونه آموزشی بسازید و به این نکته توجه داشته باشید که آموزش در شبکه‌ای که نمونه آزمایش آن کوچکتر از نمونه آموزش باشد، معمولاً کارآمدتر رخ می‌دهد.

"نمونه‌های جدا نگه داشته شده" مجموعه مستقل دیگری از داده‌های ذخیره شده است که برای ارزیابی نهایی شبکه عصبی استفاده می‌شود. خطای نمونه جدا نگه داشته شده تخمین درستی را از قابلیت پیش‌بینی مدل می‌دهد، زیرا خود این نمونه‌ها در ساخت مدل استفاده نمی‌گردند.

**Randomly assign cases based on relative number of cases (تخصیص تصادفی موارد براساس تعداد نسبی آنها)**. تعداد نسبی موارد تخصیص داده شده تصادفی را در هر یک از انواع نمونه‌ها، تعیین نمایید (آموزش، آزمایش، جدا نگه داشته شده).

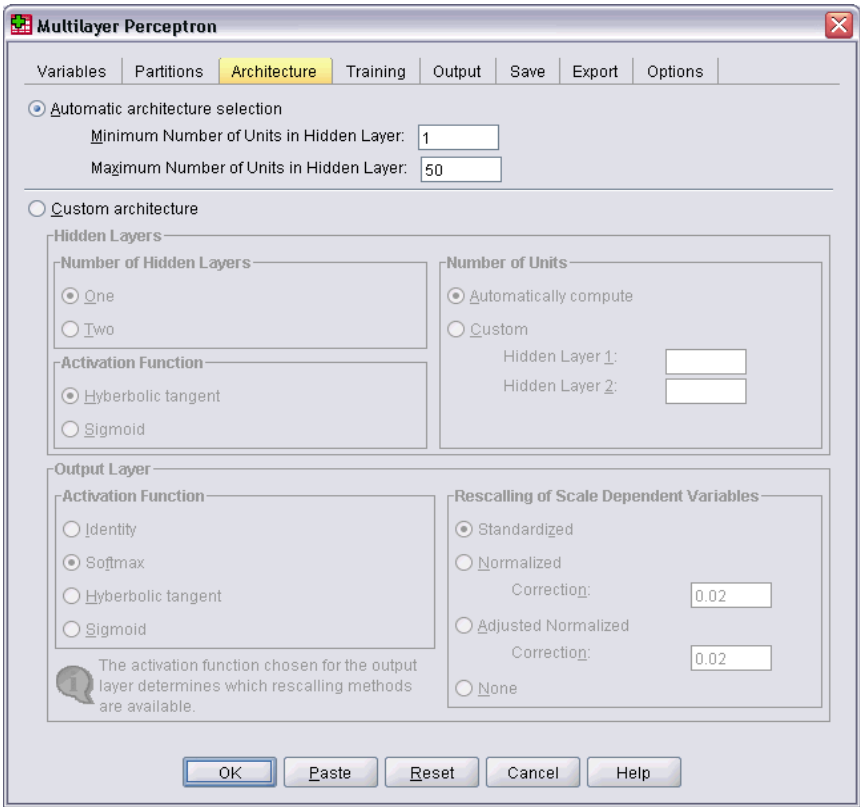
ستون ٪، درصد مواردی را که به هر نمونه براساس تعداد نسبی که شما از پیش تعیین کرده‌اید، تخصیص داده شده است را گزارش می‌دهد. به‌طور مثال، تعیین اعداد نسبی ۷ ، ۳ ، ۰ به ترتیب برای نمونه‌های آموزش، آزمایش و جدا نگه داشته شده که معادل ۷۰٪ ، ۳۰٪ ، ۰٪ می‌باشند.

یا تعیین اعداد ۱ ، ۱ ، ۲ که معادل ۵۰٪ ، ۲۵٪ ، ۲۵٪ بوده و یا همچنین تعیین ۱ ، ۱ ، ۱ که در واقع برابر با تقسیم‌بندی مجموعه داده به ۳ قسمت مساوی میان نمونه‌های آموزش، آزمایش و جدا نگه داشته شده است.

**(Use partitioning variable to assign cases.) استفاده از متغییر تفکیکی جهت تخصیص موارد.** معرف متغیری عددی است که هر یک از موارد موجود در مجموعه داده‌های فعال را به نمونه‌های آموزش، آزمایش و یا جدانگه داشته تخصیص می‌دهد. بدین‌ترتیب که مواردی که مقادیر مثبت، صفر و منفی را در متغییر تفکیکی به خود اختصاص می‌دهند، به ترتیب به گروه نمونه‌های آموزش، آزمایش و جدا از هم نگه داشته شده اختصاص می‌یابند. مواردی که هیچ مقداری را به خود اختصاص نمی دهند از آنالیزها کنار گذاشته می‌شوند.

**نکته**: استفاده از متغیر تفکیک‌کننده، به دست آمدن نتایج یکسان در اجراهای پی در پی برنامه را تضمین نمی‌کند.

## Architecture (ساختار)




*شکل ۲-۳: پرسپترون چندلایه : نوار ساختار*

نوار ساختار برای تعیین ساختار شبکه مورد استفاده قرار می‌گیرد. این فرایند می‌تواند به‌صورت خودکار "بهترین" ساختار را انتخاب کرده و یا در صورت نیاز می‌توانید یک ساختار مرسوم را خودتان تعیین کنید.

ساختار انتخاب خودکار، شبکه‌ای را با یک لایه پنهان می‌سازد. در انتخاب خودکار، ساختار بهترین تعداد واحدهای لایه پنهان را محاسبه می‌کند. انتخابگر ساختار خودکار از توابع فعال‌کننده‌ای برای لایه‌های پنهان و خروجی به‌صورت پیش‌فرض استفاده می‌کند.

انتخاب ساختار مرسوم به شما این امکان را می‌دهد که برروی لایه‌های پنهان و خروجی کنترل مناسبی داشته باشید و در مواقعی که از قبل نوع ساختاری را که می‌خواهید استفاده نمایید و یا بدان نیاز دارید را می‌دانید و بدین‌وسیله قصد بهبود ساختار خودکار را دارید، می‌تواند بسیار پر استفاده باشد.

## لایه‌های پنهان

لایه پنهان شامل گره‌ها (واحدهای) غیرقابل مشاهده شبکه است. هر واحد پنهان تابعی از حاصل جمع وزن‌دهی شده ورودی‌ها است.تابع فعال‌کننده بوده و مقادیر وزن‌دهی با الگوریتم تخمین مشخص می‌شوند. اگر شبکه لایه پنهان دومی داشته باشد، هر واحد پنهان در لایه دوم تابعی از حاصل جمع واحدهای وزن‌دهی شده در لایه اول است. تابع فعال‌کننده مشابهی در دو لایه استفاده می‌شود.

### تعداد لایه‌های پنهان

یک پرسپترون چند لایه می‌تواند یک یا دو لایه پنهان داشته باشد.

**تابع فعال‌کننده**[1]. تابع فعال‌کننده بین حاصل جمع وزن‌دهی شده واحدها در یک لایه و مقادیر واحدهای لایه بعدی ارتباط برقرار می‌کند.

- **تانژانت هیپربولیک**. شکل تابع به این صورت است:

$$\gamma(c) = \tanh(c) = \frac{(e^{c} - e^{-c})}{(e^{c} + e^{-c})}$$

این تابع مقادیر واقعی را گرفته و آنها را به مقادیری در بازه (۱،۱-) تبدیل می‌کند. وقتی از ساختار انتخابگر خودکار باشد این تابع فعال‌کننده برای تمامی واحدهای لایه‌های پنهان استفاده می‌شود.

- **سیگموئید**. شکل تابع به این صورت است:

$$\gamma(c) = \frac{1}{(1 + e^{-c})}$$

این تابع مقادیر واقعی را گرفته و آنها را به مقادیری در بازه‌ی (۱ ، ۰ ) تبدیل می‌کند.

1. Activation function

**تعداد واحدها**. تعداد واحدها در هر لایه پنهان می‌تواند مستقیم انتخاب شود و یا به‌صورت خودکار و با الگوریتم تخمین، مشخص گردد.

### لایه خروجی

لایه خروجی شامل متغیر هدف (وابسته) می‌باشد.

**تابع فعال‌کننده**. تابع فعال‌کننده بین حاصل جمع وزن‌دهی شده واحدها در یک لایه و مقادیر واحدهای لایه بعدی رابطه برقرار می‌کند.

- **همانی**[1] شکل تابع به این صورت است:

$$\gamma(c) = c$$

این تابع مقادیر حقیقی را گرفته و آنها را بدون تغییر بازمی‌گرداند. وقتی ساختار انتخابگر خودکار استفاده شود در صورت وجود متغیر وابسته به مقیاس، از این تابع فعال‌کننده برای واحدهای لایه خروجی استفاده می‌شود.

- **Softmax** . شکل تابع بدین صورت است:

$$\gamma(c_k) = \frac{\exp(c_k)}{\Sigma_j \exp(c_j)}$$

این تابع مقادیر حقیقی را گرفته و آنها را به برداری تبدیل می‌کند که المان‌های آن در بازه (۱ ، ۰ ) افتاده و حاصل جمع آنها ۱ شود. softmax تنها وقتی استفاده می‌شود که متغیرهای وابسته مطلق باشند. وقتی ساختار انتخابگر خودکار باشد، درصورتی‌که تمام متغیرهای وابسته مطلق باشند، از این تابع فعال‌کننده استفاده می‌شود.

- **تانژانت هیپربولیک**. تابع به این شکل است:

$$\gamma(c) = \tanh(c) = \frac{(e^c - e^{-c})}{(e^c + e^{-c})}$$

این تابع مقادیر واقعی را گرفته و آنها را به مقادیری در بازه (۱،۱-) تبدیل می‌کند.

- **سیگموئید**. شکل تابع به این صورت است:

$$\gamma(c) = \frac{1}{(1 + e^{-c})}$$

این تابع مقادیر واقعی را گرفته و آنها را به مقادیری در بازه‌ی (۱ ، ۰ ) تبدیل می‌کند.

---

1. Identity

(Rescaling of Scale Dependent Variables) **مقیاس‌بندی مجدد متغیرهای وابسته به مقیاس**. این کنترل‌ها تنها در حالتی که حداقل یک متغیر وابسته به مقیاس موجود باشد، انتخاب می‌شوند.

- **استاندارد شده**. میانگین از آن کم شده و بر انحراف معیار، تقسیم می‌شود.

$$\frac{x - mean}{s}$$

- **نرمال شده** . مقدار مینیمم از آن کم شده و بر دامنه تقسیم می‌شود. مقادیر نرمال شده بین صفر و یک قرار می‌گیرند. $\frac{x - min}{max - min}$ درصورتی‌که در لایه خروجی از تابع فعال‌کننده سیگموئید استفاده شود، این روش مقیاس‌بندی مجدد برای متغیرهای وابسته به مقیاس موردنیاز است. گزینه اصلاح، یک عدد کوچک (ε) را به‌عنوان اصلاحی برای فرمول مقیاس‌بندی مجدد مشخص می‌کند. علی‌الخصوص مقادیر ۰ و ۱ که وقتی x مقادیر حداقل و حداکثر خود را در فرمول نادرست اختیار کند، به‌دست می‌آیند و تعیین‌کننده بازه سیگموئید هستند اما در این بازه قرار نمی‌گیرند. فرمول اصلاح شده به این صورت است که عددی بزرگتر یا مساوی صفر می‌دهد:

$$\frac{[x - (min - \varepsilon)]}{[(max + \varepsilon) - (min - \varepsilon)]}$$

- **نرمال تنظیم شده**. مدل تنظیم شده نرمال حاصل از کم کردن مقدار حداقل و تقسیم بر دامنه می‌باشد. بدین‌صورت‌که:

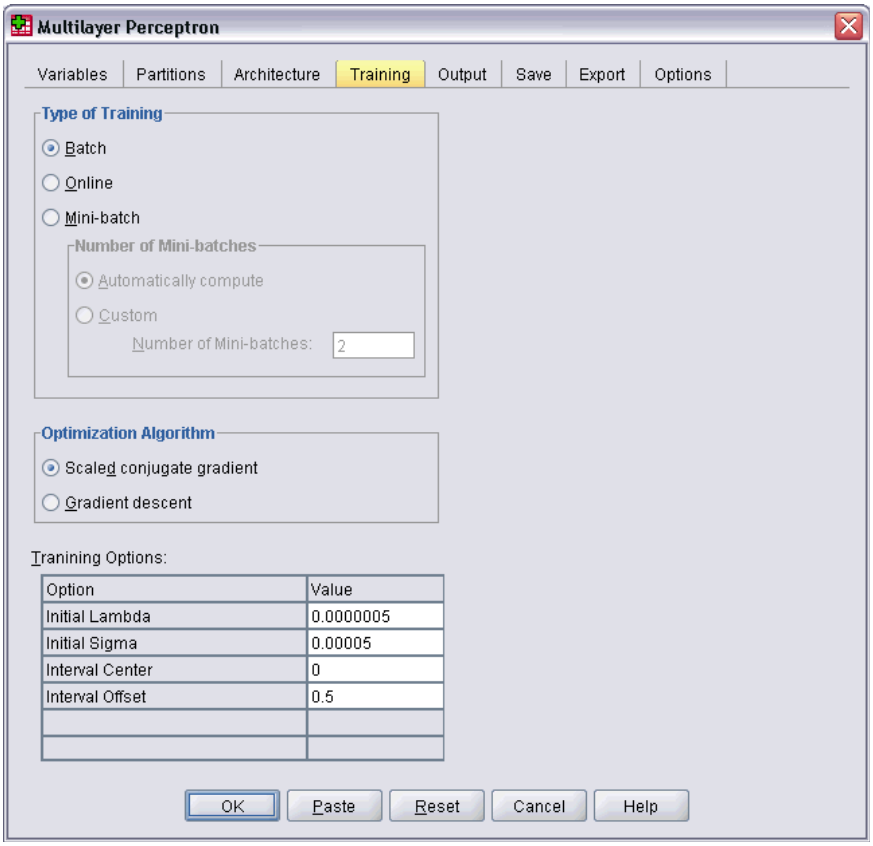
$$2 \times \frac{(x - min)}{(max - min)} - 1$$

مقادیر نرمال تنظیم شده بین ۱- و ۱ قرار می‌گیرد. درصورتی‌که تابع فعال‌کننده لایه خروجی تانژانت هیپربولیک باشد از این روش برای مقیاس‌بندی مجدد متغیرهای وابسته به مقیاس استفاده می‌شود. گزینه اصلاح، یک عدد کوچک (ε) را به‌عنوان اصلاحی برای فرمول مقیاس‌بندی مجدد مشخص می‌کند. این اصلاح قرار گرفتن تمامی متغیرهای وابسته‌ای که مجدداً مقیاس‌بندی شده‌اند را در بازه‌ی تابع فعال‌کننده تضمین می‌کند. علی‌الخصوص مقادیر ۱ و ۱- که وقتی x مقادیر حداقل و حداکثر خود را در فرمول نادرست اختیار کنند، به‌دست می‌آیند و تعیین‌کننده بازه‌ی تابع تانژانت هیپربولیک هستند، اما در این بازه قرار نمی‌گیرد. فرمول اصلاح شده بدین‌صورت است که مقادیر بزرگتر و مساوی صفر می‌دهد.

$$2 \times \frac{(x - min - \varepsilon)}{((max + \varepsilon) - (min - \varepsilon))} - 1$$

- **هیچکدام**. مقیاس‌بندی مجدد متغیرهای وابسته به مقیاس، انجام نمی‌شود.

## Training (آموزش)




*شکل ۲-۴: پرسپترون چند لایه : نوار آموزش*

نوار آموزش جهت تعیین نحوه‌ی آموزش شبکه مورد استفاده قرار می‌گیرد. نوع آموزش و الگوریتم بهینه‌سازی تعیین می‌کند که کدام گزینه‌های آموزش موجودند.

**انواع آموزش**. نوع آموزش، چگونگی پردازش داده‌های بایگانی شده را تعیین می‌کند. یکی از انواع آموزش زیر را انتخاب کنید:

- **دسته‌ای**[1]. وزن‌های سیناپسی را تنها پس از بررسی کلیه داده‌های ذخیره شده جهت آموزش به‌روز می‌رساند. بدین معنی که آموزش دسته‌ای از داده‌ها، تمام مقادیر ضبط شده در بانک اطلاعاتی آموزش را استفاده می‌کند. معمولاً آموزش دسته‌ای ترجیح داده می‌شود. زیرا این روش مستقیماً خطاهای کلی را حداقل می‌کند. در آموزش‌های دسته‌ای نیاز است که تا رسیدن به یکی از شرایط توقف وزن‌ها چندین بار به‌روزرسانی شود و بنابراین باید چندین بار بانک اطلاعات بررسی شود. این روش بیشتر برای مجموعه داده‌های کوچکتر مناسب است.
- **روی خط (Online)**. وزن‌های سیناپسی را بعد از هر داده‌ی آموزشی ذخیره شده، به روز می‌رساند. یعنی آموزش روی خط از اطلاعات مربوط به یک ذخیره در هر زمان استفاده می‌کند. آموزش روی خط به‌صورت متناوب یک ذخیزه را گرفته و وزن‌ها را به‌روزرسانی می‌کند تا زمانی که به یکی از شرایط توقف برسد. اگر تمامی ذخیره‌ها یک بار استفاده شود و به هیچ یک از شرایط توقف نرسیم، پردازش با بازیابی ذخیره‌های اطلاعات ادامه می‌یابد. آموزش روی خط در مورد مجموعه اطلاعات بزرگتر با پیش‌بینی کننده‌های مرتبط نسبت به روش دسته‌ای برتر است. درصورتی‌که ذخیره‌ها و متغیرهای متعددی موجود باشد و مقادیر آنها به یکدیگر وابسته باشند، روش آموزش روی خط سریعتر به یک پاسخ قابل قبول می‌رسد.
- **دسته‌ای**[2] – **کوچک**. بایگانی در داده‌های آموزش در گروه‌هایی تقریباً هم اندازه‌ای تقسیم شده و سپس به‌روزرسانی وزن‌های سیناپسی بعد از عبور از یک گروه انجام می‌شود. یعنی در روش دسته‌ای کوچک از اطاعات یک گروه ذخیره شده استفاده می‌شود. سپس در صورت لزوم، فرایند داده‌های گروه را بازیابی می‌کند. آموزش دسته‌ای کوچک سازشی بین "آموزش دسته‌ای" و "روی خط" است و بهترین روش برای بانک‌های اطلاعات با اندازه‌ی متوسط می‌باشد. در این روش برنامه می‌تواند به‌صورت خودکار تعداد داده‌های ضبط شده برای آموزش را مشخص کند و یا شما می‌توانید عددی بزرگتر از ۱ و کوچکتر یا مساوی حداکثر تعداد موردها را جهت ذخیره در حافظه تعیین کنید. می‌توانید حداکثر تعداد حالت‌های ذخیره شده در حافظه را در نوار option مشخص کنید.

1. Batch
2. Mini-batch

**الگوریتم بهینه‌سازی**[1]. روشی است که برای تخمین وزن‌های سیناپسی استفاده می‌شود.

- **گرادیان توام مدرج**[2]. فرضیات این روش تنها استفاده‌ی آن را برای آموزش دسته‌ای ممکن می‌سازد. بنابراین برای آموزش‌های روی خط و دسته‌ای کوچک قابل استفاده نسیت.
- **گرادیان نزولی**[3]. این روش باید برای آموزش‌های روی خط یا دسته‌ای کوچک استفاده شود. می‌توان از این روش برای آموزش دسته‌ای هم استفاده کرد.

**گزینه‌های آموزش**. گزینه‌های آموزش به شما امکان می‌دهد الگوریتم بهینه‌سازی را دقیقاً تنظیم کنید. معمولاً نیازی نیست این تنظیمات تغییر داده شوند، مگر آنکه در تخمین شبکه با مشکلی روبرو شوید. گزینه‌های آموزش برای گرادیان توام مدرج عبارتند از:

- **لاندای اولیه**[4]. مقدار اولیه ی پارامتر لاندا در الگوریتم گرادیان توام مدرج. عددی بزرگتر از صفر و کوچکتر از ۰/۰۰۰۰۰۱ را مشخص نمایید.
- **سیگمای اولیه**[5]. مقدار اولیه پارامتر سیگما در الگوریتم گرادیان توام مدرج. عددی بزرگتر از صفر و کوچکتر از ۰/۰۰۰۱ را مشخص نمایید.
- **فاصله مرکز و فاصله‌ی offset**. فاصله مرکز ($a_0$) و فاصله offset (a) ، حد فاصل $[a_0 - a , a_0 + a]$ است که در آن وقتی از بردارهای وزن annealing شبیه‌سازی شده استفاده می‌شود به‌صورت اتفاقی تولید می‌شود. annealing شبیه‌سازی شده در طی زمان استفاده از الگوریتم بیهنه‌سازی، برای خروج از مینیمم موضعی و با هدف پیدا کردن مینیمم کلی، حاصل می‌شود. این روش در وزن‌دهی اولیه و انتخاب ساختار خودکار استفاده می‌شود. عددی برای فاصله مرکز و عددی بزرگتر از صفر برای فاصله‌ی offset مشخص کنید.

گزینه‌های آموزش برای گرادیان نزولی عبارتند از:

- **سرعت آموزش اولیه** . مقدار اولیه سرعت آموزش برای الگوریتم گرادیان نزولی .
- سرعت آموزش بالاتربه معنی آموزش سریعتر شبکه بوده که می‌تواند با ناپایداری آن همراه شود. عددی بزرگتر از صفر را مشخص کنید.

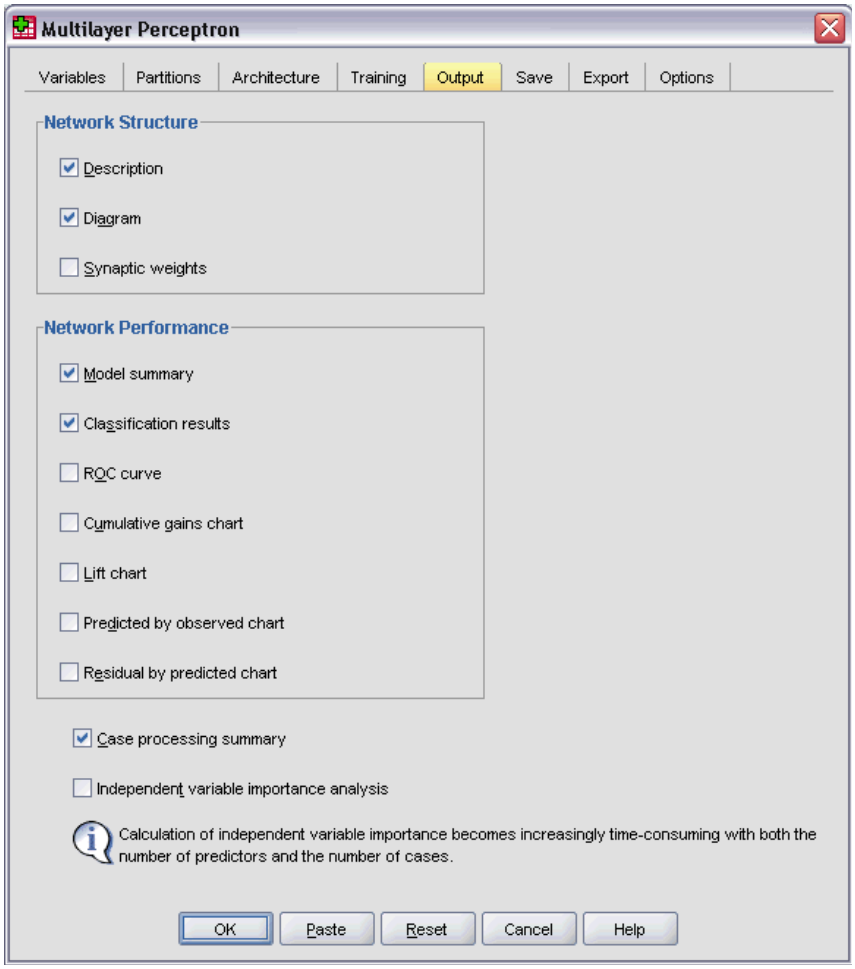
---

1. Optimization algorithm
2. Scaled conjugate gradient
3. Gradient discent
4. Initial lambda
5. Initial sigma

- **حد پایین سرعت آموزش**[1] . این تنظیم تنها برای آموزش روی خط و دسته‌ای کوچک انجام می‌شود. عددی بزرگتر از صفر و کوچکتر از سرعت اولیه آموزش مشخص کنید.
- **اندازه حرکت**[2] . مقدار اولیه پارامتر اندازه حرکت برای الگوریتم گرادیان نزولی.
- واژه سرعت حرکت به جلوگیری از عدم پایداری ناشی از سرعت آموزش خیلی بالا کمک می‌کند. عددی بزرگتر از صفر مشخص کنید.
- **کاهش سرعت آموزش در مبداء**[3] . وقتی کاهش گرادیان با آموزش روی خط یا دسته‌ای کوچک حاصل می‌گردد، باید تعداد مبدأها (P) یا اطلاعات منتقل شده از نمونه آموزشی جهت کاهش سرعت اولیه آموزش تا حدود پایینی سرعت آموزش کم شود. این کار به شما امکان می‌دهد که ضریب تنزل $\beta = (1/pK)\times\ln(\eta_0/\eta_{low})$ را که در آن $\eta_0$ سرعت اولیه آموزش، $\eta_{low}$ حدود پایین سرعت آموزش و K تعداد کل دسته‌های کوچک (یا تعداد آموزش‌های ضبط شده برای آموزش روی خط) است را کنترل نمایید. عدد صحیح بزرگتر از صفر تعیین کنید.

1. Initial Learning Rate
2. momentum
3. Learning rate reduction, in Epochs

## OutPut (خروجی)




*شکل ۲-۵ : پرسپترون چند لایه : نوار خروجی*

**ساختار شبکه**. خلاصه اطلاعات شبکه عصبی را نشان می‌دهد.

- **توصیف**[1]. اطلاعاتی مربوط به شبکه عصبی را که شامل متغیرهای وابسته، تعداد واحدهای ورودی و خروجی، تعداد واحدها و لایه‌های پنهان و توابع فعال‌کننده هستند را، نشان می‌دهد.

1. Description

- **دیاگرام**. دیاگرام شبکه را به‌صورت یک نمودار غیرقابل تغییر نشان می‌دهد. قابل ذکر است هر چه تعداد متغیر کمکی و ضریب سطوح افزایش یابد، تفسیر دیاگرام دشوارتر می‌شود.
- **وزن‌های سیناپسی**. ضریب‌هایی که برای نشان دادن رابطه بین واحدهای یک لایه و لایه بعد تخمین‌زده شده اند را نشان می‌دهد. حتی اگر بانک اطلاعات فعال به داده‌های آموزش، آزمایش و جدانگه داشته شده تقسیم شود، وزن‌های سیناپسی براساس داده‌های آموزش تعیین می‌شوند شایان ذکر است که تعداد وزن‌های سیناپسی می‌تواند زیاد باشد و معمولاً از این وزن‌ها برای تفسیر شبکه عصبی استفاده نمی‌شود.

**عملکرد شبکه**[1]. نتایجی را که جهت نشان دادن "خوب بودن " مدل مورد استفاده قرار می‌گیرند نمایش می‌دهد.

**نکته**: نمودارهای این گروه براساس مجموعه‌ای ترکیبی از نمونه‌های آموزش و آزمایش ترسیم می‌شوند. با توجه به این نکته که در صورت عدم وجود نمونه آزمایش تنها براساس نمونه آموزش رسم می‌گردند.

- **خلاصه مدل**[2]. خلاصه‌ای از نتایج شبکه عصبی را به شکل کامل و به تفکیک ارائه می‌دهد از جمله این موارد می‌توان به خطاها، خطای نسبی (درصد پیش‌بینی نادرست)، قوانین توقف جهت متوقف کردن آموزش و زمان آموزش، اشاره نمود.

زمانی که برای تابع فعال‌کننده لایه خروجی، از توابعی همچون، تابع همانی، سیگموئید یا تانژانت هیبربولیک استفاده نماییم، از خطای حاصل جمع مربعات استفاده می‌شود و چنانچه تابع فعال‌کننده لایه خروجی softmax باشد، خطای cross enthropy مورد استفاده قرار می‌گیرد.

خطای نسبی یا درصد پیش‌بینی نادرست بسته به سطوح اندازه‌گیری متغیر وابسته نمایش داده می‌شود. اگر یکی از متغیرهای وابسته سطح اندازه‌گیری مقیاس‌بندی شده داشته باشد، میانگین خطای نسبی کلی (مربوط به مدل میانگین) نمایش داده می‌شود. اگر تمامی متغیرهای وابسته مطلق باشند، میانگین درصد پیش‌بینی نادرست نمایش داده می‌شود. خطاهای نسبی یا درصدهای پیش‌بینی نادرست برای تک‌تک متغیرهای وابسته نیز مشخص می‌گردند.

1. Network Performance
2. Model Summary

- **نتایج طبقه‌بندی** . یک جدول طبقه‌بندی برای هر متغیر وابسته مطلق به‌صورت جزئی و کلی ارائه می‌دهد. هر جدول تعداد مواردی را که در هر یک از دسته‌بندی‌های مربوط به متغیر وابسته، به‌صورت درست یا نادرست طبقه‌بندی شده‌اند را نشان می‌دهد. درصد کل حالت‌هایی که درست طبقه‌بندی شده‌اند نیز گزارش می‌شود.
- **منحنی ROC** . برای هر متغیر وابسته مطلق منحنی ROC نمایش داده می‌شود. همچنین جدولی که نمایانگر سطح زیر هر منحنی است ارائه می‌گردد. زمانی که یک متغیر وابسته داده شد، نمودار ROC برای هردسته یک منحنی نمایش می‌دهد. چنانچه متغیر وابسته، ۲ دسته داشته باشد، هر منحنی با دسته‌ی مربوطه به شکل مثبت در مقابل دسته‌ی دیگر فرض می‌شود. اگر متغیر وابسته بیش از ۲ دسته داشته باشد، باز هم هر منحنی با دسته‌ی مربوطه به شکل مثبت فرض می‌گردد با این تفاوت که این بار در مقابل مجموعه‌ای یکپارچه شده از سایر دسته‌ها قرار می‌گیرد.
- **نمودار بهره تجمعی**[1]. یک نمودار بهره تجمعی برای هر متغیر وابسته مطلق نمایش داده می‌شود. در اینجا نیز مشابه منحنی‌های ROC ، برای هر دسته‌ی متغیر وابسته یک منحنی ارائه می‌شود.
- **نمودار lift**[2]. یک نمودار lift برای هر متغیر وابسته‌ی مطلق نمایش داده می‌شود. در اینجا نیز مشابه منحنی ROC ، برای هر دسته‌ی متغیر وابسته یک منحنی ارائه می‌شود.
- **نمودار پیش‌بینی شده در مقابل مشاهده شده**[3] . یک نمودار پیش‌بینی براساس مشاهده برای هر متغیر وابسته ارائه می‌شود. برای متغیرهای وابسته مطلق، نمودار میله‌ای تجمعی برای نمایش شبه احتمال‌های پیش‌بینی شده در هر دسته استفاده شده است. برای متغیرهای وابسته به مقیاس یک نمودار پراکندگی نیز ارائه می‌گردد.
- **منحنی باقی مانده در مقابل پیش‌بینی شده**[4] . یک منحنی مقادیر باقی مانده براساس پیش‌بینی، برای هر متغیر وابسته به مقیاس ارائه می‌دهد. هیچ ساختار مشخصی نباید بین

1. Cumulative gains chart.
2. Lift curve
3. Predicted by observed chart
4. Residual by predicted chart.

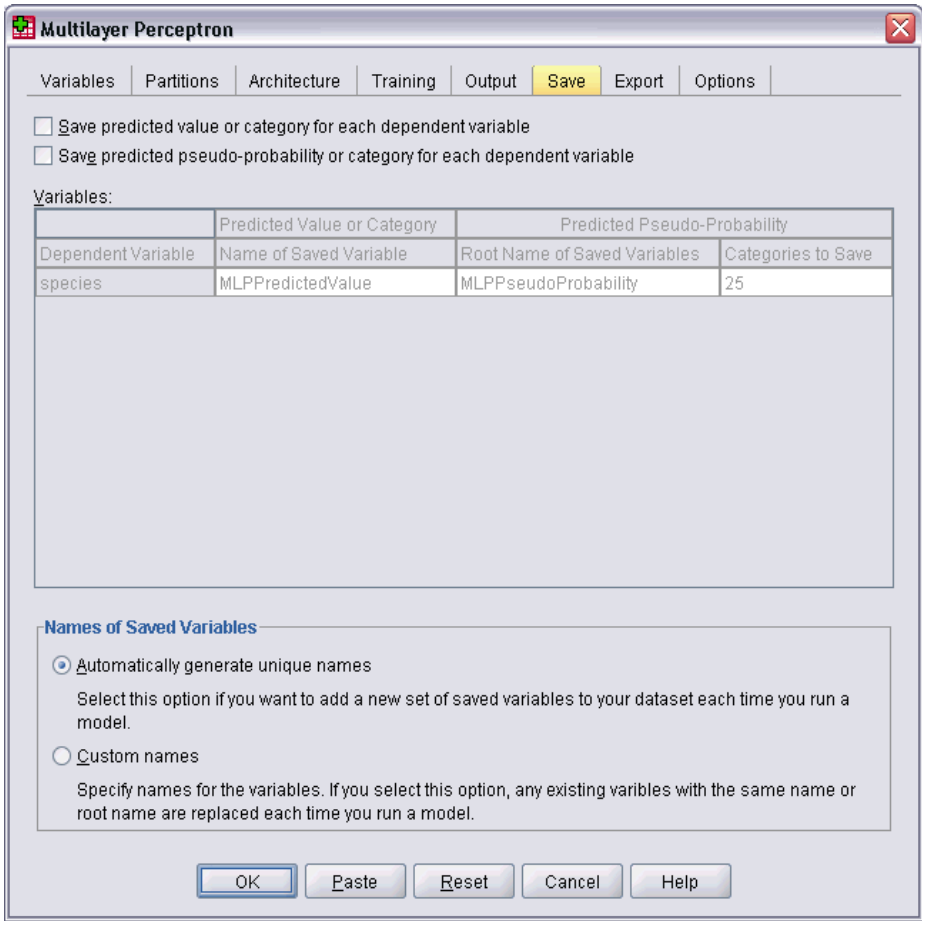
مقادیر باقی مانده و پیش‌بینی شده وجود داشته باشد. این نمودار تنها برای متغیرهای وابسته به مقیاس ترسیم می‌شود.

**خلاصه پردازش حالت** [1]. جدول خلاصه پردازش‌های انجام شده، تعداد مواردی را که در آنالیزها استفاده شده‌اند و همچنین تعداد مواردی که از آنالیزها خارج شده‌اند را به‌صورت خلاصه نمایش می‌دهد.

- **تحلیل اهمیت متغیر مستقل** [2]. آنالیز حساسی است که اهمیت هر پیش‌بینی‌کننده در تعیین شبکه عصبی را محاسبه می‌کند. تحلیل ممکن است برپایه نمونه‌های آموزش و آزمایش تلفیق شده یا در صورت عدم وجود نمونه آزمایش تنها روی نمونه آموزش انجام شود. در نهایت یک جدول و یک نمودار که نشان‌دهنده اهمیت و اهمیت نرمال شده هر پیش‌بینی‌کننده است، ارائه می‌شود. شایان ذکر است که تحلیل حساسیت در صورت وجود تعداد زیاد پیش‌بینی‌کننده‌ها یا حالات گران قیمت و زمان‌بر است.

1. Case processing summary
2. Independent variable importance analysis

## ذخیره (save)




*شکل ۲-۶: پرسپترون چندلایه : نوار ذخیره*

نوار ذخیره جهت ذخیره پیش‌بینی‌ها به‌صورت متغیر در بانک اطلاعات مورد استفاده قرار می‌گیرد.

- **مقادیر پیش‌بینی شده** (**Save predicted value or category for each dependent variable**) **یا دسته هر متغیر وابسته را ذخیره کن**. این گزینه مقادیر پیش‌بینی شده هر یک از متغیرهای وابسته به مقیاس و همچنین دسته پیش‌بینی شده برای متغیرهای وابسته مطلق را ذخیره می‌کند.

- **شبه احتمال پیش‌بینی شده یا دسته هر متغیر وابسته را ذخیره کن** (**Save predicted pseudo-probability or category for each dependent variable**). این گزینه شبه احتمال پیش‌بینی شده برای متغیر وابسته مطلق را ذخیره می‌کند. یک متغیر جداگانه برای هر n دسته‌ی اول، که n در ستون (categories to save) تعیین می‌شود، ذخیره خواهد شد.

**نام متغیرهای ذخیره شده** . تولید نام به‌صورت خودکار تضمین می‌کند که شما تمامی کار خود را حفظ کرده‌اید.

نام‌های مرسوم به شما امکان می‌دهد که نتایج را از اجراهای قبلی بدون پاک کردن متغیرهای ذخیره شده در ویرایشگر داده حذف یا جایگزین کنید.

## احتمال‌ها وشبه احتمال‌ها[1]

متغیرهای وابسته با تابع فعال‌ساز softmax و خطای cross-enthropy یک مقدار پیش‌بینی شده برای هر دسته دارند، که این مقدار پیش‌بینی شده احتمال متعلق بودن یک مورد مشخص به یک دسته مشخص است. متغییرهای وابسته مطلق با مجموع مربعات خطا نیز، یک مقدار پیش‌بینی شده برای هر دسته دارند اما این مقدار پیش‌بینی شده نمی‌تواند به‌عنوان احتمال تعبیر شود. برنامه، این مقادیر شبه احتمال پیش‌بینی شده را نیز حتی اگر کمتر از صفر یا بزرگتر از ۱ بوده و یا حاصل جمع مقادیر یک متغیر وابسته خاص، ۱ نباشد، ذخیره می‌کند.

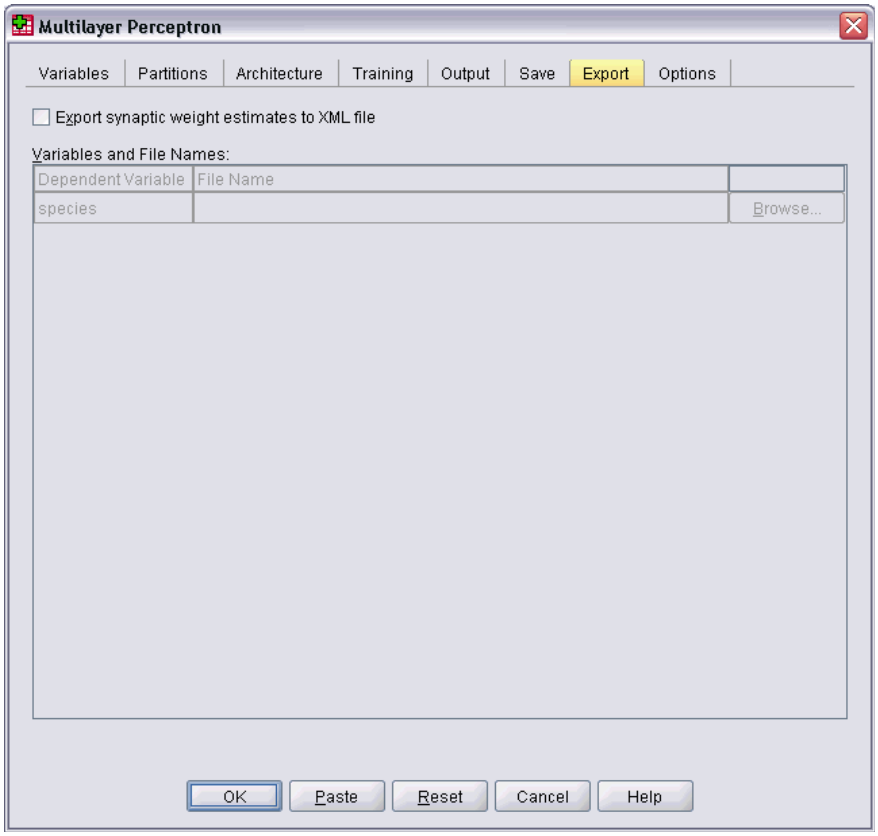
ROC ، بهره تجمعی و نمودار Lift براساس شبه احتمال‌ها ساخته شده‌اند. در حالتی که هر کدام از شبه احتمال‌ها کمتر از صفر یا بزرگتر از ۱ باشند، یا حاصل جمع برای یک متغیر ۱ نباشند، این مقادیر مقیاس‌بندی مجدد می‌شوند تا در بازه‌ی صفر تا ۱ قرار گرفته و حاصل جمع‌شان ۱ شود. شبه احتمال با تقسیم شدن به حاصل جمعشان مقداردهی مجدد می‌شوند. برای مثال، وقتی برای یک متغیر وابسته سه دسته‌ای مقادیر شبه احتمال پیش‌بینی شده ۰/۵ ، ۰/۶ و ۰/۴ باشد، هر کدام از اعداد بر ۱/۵ تقسیم می‌شود تا ۰/۳۳ ، ۰/۴ و ۰/۲۷ حاصل شود.

اگر هر یک از شبه احتمال‌ها منفی باشد، قدر مطلق کمترین عدد قبل از مقیاس‌بندی مجدد به همه مقادیر اضافه می‌شود. برای مثال اگر شبه احتمال‌ها ۰/۳- ، ۰/۵- و ۱/۳ باشند، ابتدا ۰/۳

1. Probabilities and Pseudo-Probabilities

به هر مقدار اضافه می‌شود تا ۰ ، ۰/۸ و ۱/۶ به دست آید. سپس هر مقدار به حاصل جمع ۲/۴ است تا به ۰ ، ۰/۳۳ و ۰/۶۷ برسیم.

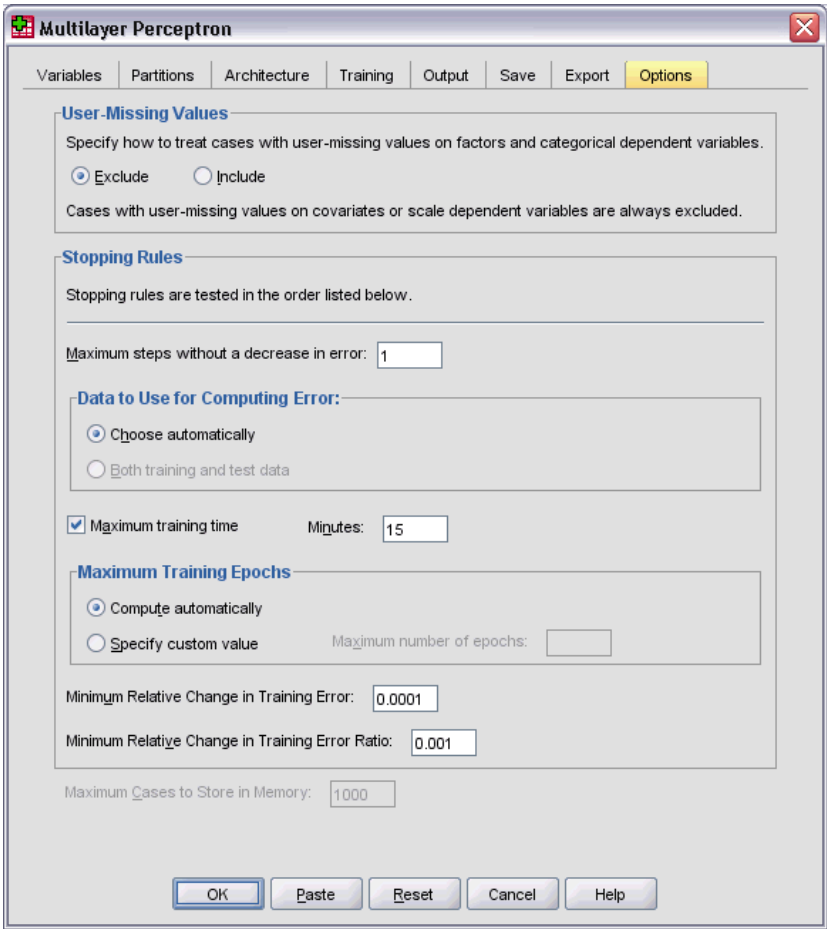
## Export (صدور)




*شکل ۲-۷ : پرسپترون چندلایه : نوار صدور*

نوار Export جهت ذخیره تخمین وزن‌های سیناپسی هر متغیر وابسته برروی یک فایل XML(PMML) استفاده می‌شود.

## Options (گزینه‌ها)




*شکل ۲-۸: پرسپترون چندلایه : نوار گزینه ها*

**مقادیر از دست رفته کاربر**[1] . فاکتورها باید مقادیر قابل دسترسی برای یک مورد داشته باشند تا در تحلیل‌ها وارد شوند. این کنترل‌ها به شما امکان می‌دهد تصمیم بگیرید که مقادیر از دست رفته کاربر در فاکتورها و متغیرهای وابسته مطلق موجود باشند یا خیر.

**قوانین متوقف کننده**[2] . اینها قوانینی هستند که زمان توقف آموزش شبکه عصبی را مشخص می‌کنند. آموزش حداقل با عبور یک داده ادامه می‌یابد.آموزش با توجه به ضوابطی که از روی

1. User-Missing Values.
2. Stopping Rules

یک دستورالعمل مدون کنترل شده برداشت شده است، متوقف می‌شود. در تعیین قوانین متوقف‌کننده فرآیند باید این نکته را در نظر داشت که در روش‌های آموزش روی خط و دسته‌ای کوچک، هر یک قدم برابر با عبور یک مورد و در روش دسته‌ای برابر با تکرارپذیری خواهد بود.

**حداکثر مراحل بدون کاهش در میزان خطا** [1]. تعداد مراحل مجاز قبل از بررسی کاهش در خطا. اگر بعد از تعداد مشخصی مرحله، کاهشی در خطا مشاهده نشود، آموزش متوقف خواهد شد. عددی بزرگتر از صفر تعیین کنید. همچنین می‌توانید تعیین کنید از کدام نمونه داده جهت محاسبه خطا استفاده شود. در صورت وجود نمونه آزمایش به‌صورت خودکار انتخاب می‌شود و در غیر این صورت از نمونه آموزش استفاده می‌گردد. شایان ذکر است که آموزش دسته‌ای کاهش میزان خطای نمونه آموزش را بعد از هر بار انتقال داده را تضمین می‌کند. بنابراین، تنها وقتی نمونه آزمایش موجود باشد از این روش برای آموزش دسته‌ای استفاده می‌شود. داده‌های آموزش و آزمایش هر دو میزان خطا را برای هر کدام از این نمونه ها بررسی می کنند.این گزینه تنها وقتی که نمونه آزمایش موجود باشد مورد استفاده قرار می‌گیرد.

- **حداکثر زمان آموزش**[2]. می‌توان حداکثر زمان اجرای الگوریتم را مشخص نمود. عددی بزرگتر از صفر تعیین کنید.
- **حداکثر مبدأهای آموزش**[3]. حداکثر تعداد مجاز مبدأها (انتقال داده) . اگر از حداکثر تعداد مبدأها تجاوز کنیم، آموزش متوقف می شود. عدد صحیحی بزرگتر از صفر مشخص کنید.
- **حداقل تغییرات نسبی در خطای آموزش** [4]. درصورتی‌که تغییر نسبی در خطای آموزشی در مقایسه با مرحله قبل از مقیاس کمتر شود، عدد صحیحی بزرگتر از صفر مشخص کنید. برای آموزش روی خط و دسته‌ای کوچک اگر تنها از داده‌های آزمایش برای محاسبه خطا استفاده شود ، این مقیاس نادیده گرفته می‌شود.

---

1. Maximum steps without a decrease in error
2. Maximum training time
3. Maximum Training Epochs
4. Minimum Relative Change in training error

- **حداقل تغییر نسبی در نسبت خطای آموزش**[1] . درصورتی‌که نسبت خطای آموزش به خطای مدل تهی کمتر از مقیاس معینی باشد، آموزش متوقف می‌شود. مدل تهی مقدار میانگین تمام متغیرهای وابسته را پیش‌بینی می‌کند. عددی بزرگتر از صفر تعیین کنید. بر آموزش روی خط و دسته‌ای کوچک اگر تنها از داده‌های آموزش برای محاسبه خطا استفاده شود. این مقیاس نادیده گرفته می‌شود.

**حداکثر حالاتی که در حافظه ذخیره می‌شوند**[2]. تنظیمات زیر در الگوریتم پرسپترون چند لایه با این گزینه کنترل می‌شود. عددی بزرگتر از ۱ مشخص کنید.

- در انتخاب ساختار خودکار، اندازه نمونه‌ای که جهت تعیین ساختار شبکه استفاده می‌شود min(memsize،۱۰۰۰) است ؛ که memsize حداکثر تعداد حالات ذخیره شده در حافظه می‌باشد.

در آموزش دسته‌ای کوچک با محاسبه خودکار تعداد دسته‌های کوچک، تعداد دسته‌های کوچک برابر خواهد بود با $\min(\max(M/10,2),memsize)$ که در اینجا M تعداد حالات در نمونه آموزشی می‌باشد.

1. Minimum relative change in training error Ratio
2. Meximum Cases to store in memory

# بخش سوم

## تابع شعاع مدار

روش تابع شعاع مدار (RBF) یک مدل پیش‌بینی‌کننده برای یک یا چند متغیر وابسته (هدف) براساس مقادیر متغیرهای پیش‌بینی‌کننده ، فراهم می‌کند.

**مثال**. یک سرویس‌دهنده ارتباط از راه دور، مشتریانش را براساس نوع استفاده از خدمات به ۴ دسته طبقه‌بندی می‌کند. یک شبکه RBF با استفاده از اطلاعات آمارگیری جهت پیش‌بینی عضویت در گروه‌ها، به شرکت این امکان را می‌دهد که پیشنهاد مناسب برای مشتریان بعدی آماده کند.

## متغیرهای وابسته

متغیرهای وابسته می‌توانند به این صورت باشند:

- **اسمی**. در این نوع مقیاس برای اندازه‌گیری متغیر از اسامی، حروف و... استفاده می‌شود مانند اسم افراد، اسم شرکت‌ها و...
- **مقیاس ترتیبی**. در این نوع مقیاس ترتیبی در سطوح متغیر وجود دارد مانند طیف پنج گزینه‌ای خیلی کم، کم، متوسط، زیاد و خیلی زیاد برای رضایت کارکنان یا طیف سه گزینه‌ای درجه ۱ ، ۲ و ۳ برای کیفیت محصول.
- **مقیاس فاصله‌ای**. مقیاس کمی است که نقطه صفر آن قراردادی می‌باشد و فاصله بین دو واحد متوالی آن مقدار ثابتی است. مقیاس‌های سانتی‌گراد و فارنهایت نمونه‌هایی از این مقیاس می‌باشند.
- **مقیاس نسبی (نسبتی)**. مشابه مقیاس فاصله‌ای است با این تفاوت که نقطه صفر آن واقعی می‌باشد مانند مقیاس های سانتی‌متر و اینچ.

در این مراحل فرض شده است که سطح مناسب اندازه‌گیری برای تمام متغیرهای وابسته نسبت داده شده است. به هر حال شما می‌توانید موقتاً سطوح اندازه‌گیری برای یک متغیر را با کلیک راست روی متغیر در لیست آن و انتخاب یک سطح اندازه‌گیری از فهرست تغییر دهید.

یک نشانه گر درکنار هر متغیر در لیست متغیر سطح اندازه‌گیری و نوع داده آن را مشخص می‌کند.

| Measurement Level | Data Type | | | |
|---|---|---|---|---|
| | Numeric | String | Date | Time |
| Scale |  | n/a |  |  |
| Ordinal |  |  |  |  |
| Nominal |  |  |  |  |

**متغیر پیش‌بینی‌کننده**. پیش‌بینی‌کننده‌ها می‌توانند به‌صورت مطلق یا نسبی تعیین شوند.

**کدگذاری متغیر مطلق**. فرایند موقتاً متغیرهای وابسته و پیش‌بینی‌کننده مطلق را با استفاده از کدهای ۱ از c تا اتمام مراحل، کدگذاری مجدد می‌کند. اگر c دسته از یک متغیر وجود داشته باشد، به‌عنوان بردار c ذخیره می‌شود، اولین نوع بردار (1,0,…,0) بعدی (0,1,0,…,0) و بالاخره آخرین نوع (0,0,…,1) است.

این مدل کدگذاری تعداد وزن‌های سیناپسی را افزایش داده و باعث آموزش آهسته‌تر می‌شود. به هر حال هرچه روش‌های کدگذاری فشرده‌تر باشند معمولاً به انطباق ضعیف‌تری با شبکه عصبی منجر خواهد شد. اگر شبکه‌ی آموزشی شما پیشرفت بسیار آهسته‌ای دارد سعی کنید تعداد پیش‌بینی‌کننده‌های مطلق خود را به‌وسیله ترکیب کردن با دسته‌های مشابه و یا حذف دسته‌های خیلی نادر، کاهش دهید.

حتی اگر یک نمونه آزمایش یا "نمونه جدا نگه داشته شده" مشخص شده باشد، همه‌ی کدگذاری ۱ از c برپایه داده آموزش است.اگر نمونه جدا نگه داشته شده یا نمونه آزمایش شامل مواردی باشد که دسته‌های پیش‌بینی‌کننده آن در داده آموزش ارائه نشده است، از آنها در برنامه یا امتیازدهی استفاده نخواهند شد. چنانچه نمونه‌های آزمایش یا جدا از هم نگه داشته شده شامل حالتی از دسته‌های متغیر وابسته باشد که در داده‌های آموزش آورده نشده است، این حالت‌ها در فرایند استفاده نشده اما ممکن است امتیازدهی شوند.

**مقیاس‌بندی مجدد**. متغیرها و متغیرهای کمکی وابسته به مقیاس جهت بهبود آموزش شبکه به‌صورت قراردادی مقیاس‌بندی مجدد می‌شوند. حتی به‌صورت وجود نمونه‌های آزمایش و جدا نگه داشته شده نیز مقیاس‌بندی‌های مجدد براساس داده های آموزش انجام می‌شود.

بدین‌معنی که براساس نوع مقیاس‌بندی مجدد، میانگین، انحراف معیار، مقدار حداقل یا حداکثر یک متغیر کمکی یا متغیر وابسته تنها با استفاده از داده‌های آموزش محاسبه می‌شوند. تبعیت کردن این متغیرهای کمکی یا وابسته از یک توزیع مشابه در نمونه‌های آموزش، آزمایش و جدا نگه داشته شده حین تفکیک یک متغیر خاص مهم است.

**وزن‌های فراوانی**. فراوانی وزن در این روش نادیده گرفته می‌شود.

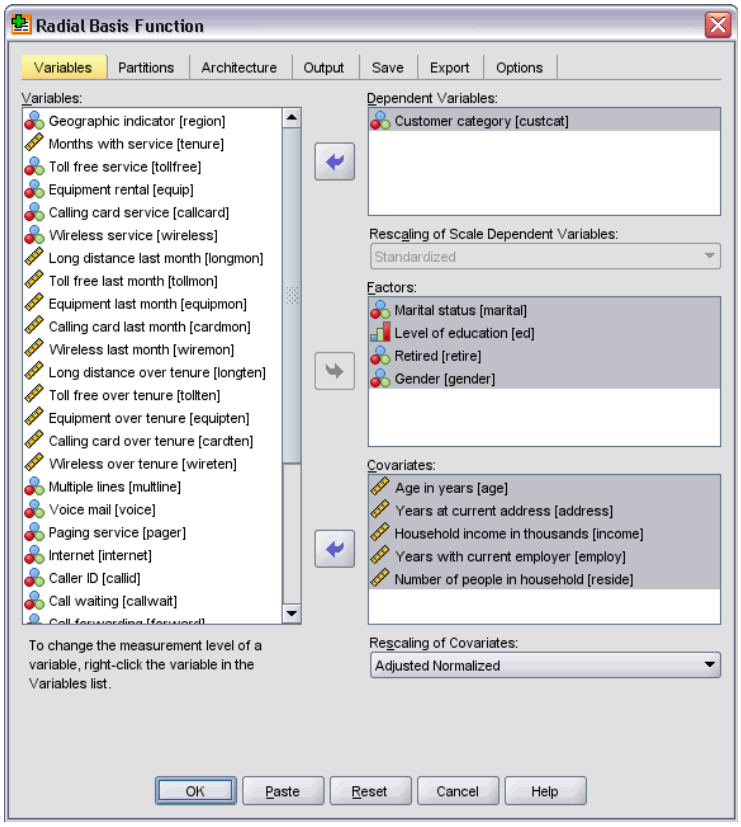
**تکرار در نتایج**. اگر شما می‌خواهید دقیقاً نتایج را تکرار کنید، از همان مقدار اولیه برای تولید ارقام تصادفی، همان داده و همان متغیر استفاده کنید. علاوه براین از همان تنظیمات استفاده کنید. جزئیات بیشتر این مسئه به شرح زیر است:

- **تولید اعداد تصادفی**. این برنامه از تولید تصادفی اعداد در تخصیص تصادفی قسمت‌ها استفاده می‌کند. برای تولید دوباره‌ی همان نتایج تصادفی به‌دست آمده در آینده، از همان مقدار اولیه قبل از هر بار اجرای روش RBFبرای تولید اعداد تصادفی استفاده کنید.
- **مرتبه حالت**. نتایج به علت این که از الگوریتم خوشه دو مرحله‌ای برای تعیین تابع شعاع مدار استفاده می‌شود، به مرتبه حالت وابسته‌اند.
- برای حداقل کردن اثرات مرتبه‌ای به‌طور تصادفی حالات را مرتب کنید. جهت بررسی پایداری مراحل ارایه شده، نیاز دارید راه‌حل‌های متفاوت با طبقه‌بندی‌های مختلف تصادفی داشته باشید. در شرایطی که اندازه فایل خیلی بزرگ است، چندین اجرا برروی یک نمونه از حالت‌هایی که در مراتب تصادفی ذخیره شده‌اند، کافی است.

## ساخت یک شبکه تابع شعاع مدار

از منو انتخاب کنید:

analyze→neural network→Radial Basis Function




*شکل ۲-۹ : تابع شعاع مدار: نوار متغیر*

- یک متغیر وابسته انتخاب کنید
- یک ضریب یا متغیر کمکی انتخاب کنید
- شما می‌توانید روی نوار variable به‌طور اختیاری روش مقیاس‌بندی مجدد متغیرهای کمکی را تغییر دهید.

گزینه‌ها عبارتند از:

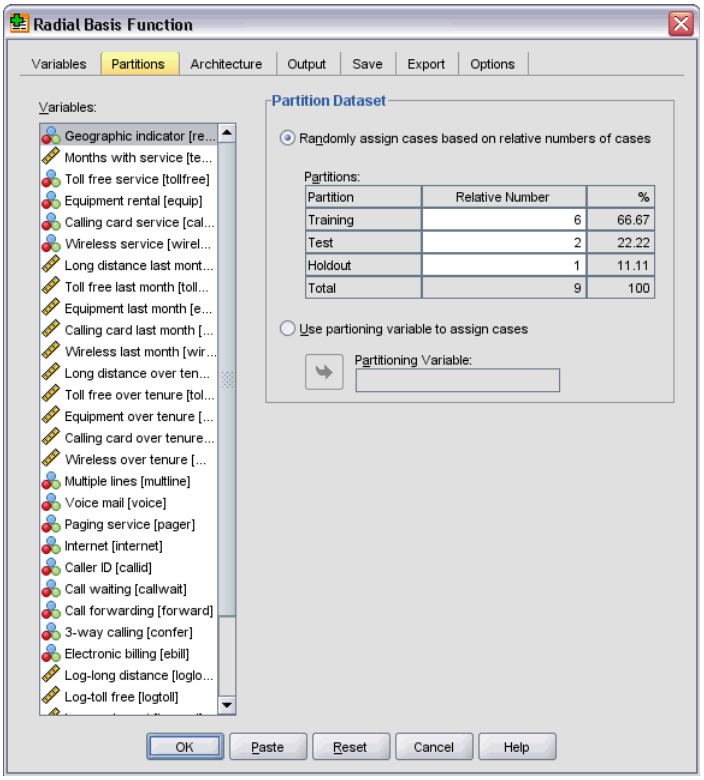
- استاندارد شده: میانگین از آن کم شده و بر انحراف معیار تقسیم می‌شود. $\frac{x-mean}{s}$
- نرمال شده: مقدار حداقل از آن کم شده و بر دامنه تقسیم می‌شود. $\frac{x-min}{max-min}$

  مقدار آن بین ۰ و ۱ است.
- نرمال تنظیم شده : کم کردن مقدار حداقل و تقسیم بر دامنه $\frac{2(x-min)}{max-min}-1$

مقدار آن بین ۱- و ۱ است.

- هیچکدام: بدون مقیاس‌بندی مجدد متغیرهای کمکی

## Partiotions (تفکیک کردن)




*شکل ۲-۱۰: تابع شعاع مدار: نوار Partitions*

**تفکیک مجموعه داده**: این قسمت روش تفکیک مؤثر مجموع داده را به نمونه‌های آزمایش، جدا نگه داشته شده و نمونه‌های آموزش تعیین می‌کند.

"نمونه آزمایش" شامل داده‌های ثبت شده مورد استفاده در آموزش شبکه عصبی می‌باشد. کسری از موارد در مجموعه داده باید به نمونه آموزش جهت به‌دست آوردن مدل، تخصیص داده شوند. نمونه آزمایش یک مجموعه مستقل داده‌های ذخیره شده برای پیدا کردن خطا حین آموزش به‌منظور جلوگیری از آزموش زیاد است.

به شدت توصیه شده است که یک نمونه آموزشی بسازید و به این نکته توجه داشته باشید که آموزش در شبکه‌ای که نمونه آزمایش آن کوچکتر از نمونه آموزش باشد، معمولاً کارآمدتر رخ می‌دهد.

"نمونه‌های جدا نگه داشته شده" مجموعه مستقل دیگری از داده‌های ذخیره شده است که برای ارزیابی نهایی شبکه عصبی استفاده می‌شود. خطای نمونه جدا نگه داشته شده تخمین درستی از قابلیت پیش‌بینی مدل می‌دهد، زیرا این نمونه‌ها در ساخت مدل استفاده نمی‌شوند.

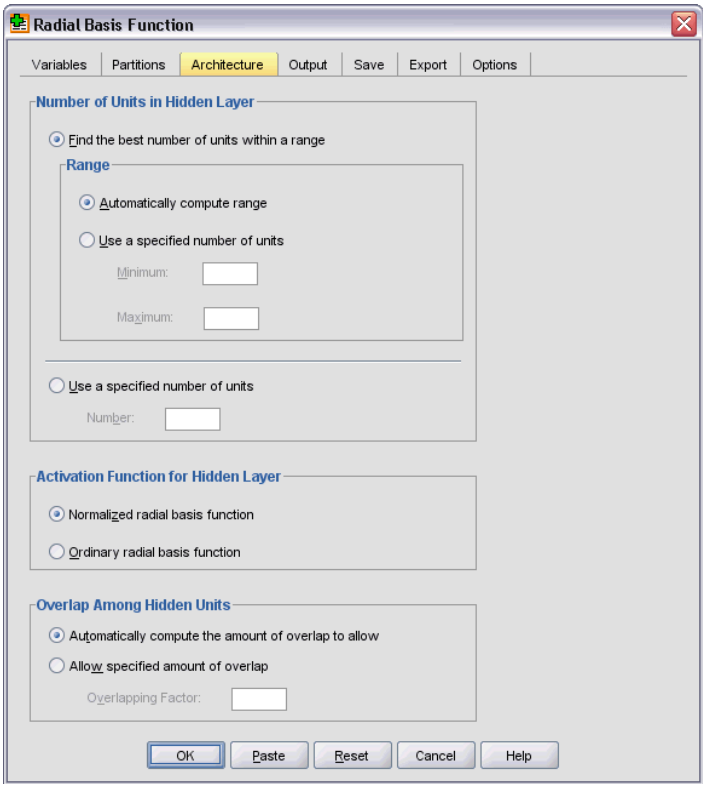
- Randomly assign cases based on relative number of cases (تخصیص تصادفی حالات براساس تعداد نسبی حالت‌ها). تعداد نسبی حالات تصادفی تخصیص داده شده هر نمونه را تعیین کنید (آموزش، آزمایش، جدا نگه داشته شده).

ستون ٪ درصد حالاتی که به هر نمونه تخصیص داده شده‌اند براساس تعداد نسبی که شما تعیین کردید گزارش می‌دهد. به‌طور مثال، تعیین اعداد نسبی ۷ ، ۳ ، ۰ برای آموزش، آزمایش و جدا نگه داشته شده معادل ۷۰٪ ، ۳۰٪ ، ۰٪ است.

یا تعیین اعداد ۱ ، ۱ ، ۲ معادل ۵۰٪ ، ۲۵٪ ، ۲۵٪ است. ۱ ، ۱ ، ۱ برابر با تقسیم‌بندی مجموعه داده به ۳ قسمت مساوی میان نمونه‌های آموزش، آزمایش و جدا نگه داشته شده می‌باشد.

- **استفاده از تفکیک متغیرها جهت تخصیص حالات** . متغیر عددی که هر حالت از مجموعه داده فعال را به نمونه‌های آموزش، آزمایش یا جدانگه داشته تخصیص می‌دهد تعیین کنید. حالاتی با مقدار مثبت متغیرها به نمونه آموزش، مقدار صفر به نمونه آزمایش و مقدار منفی به نمونه جدا نگه داشته شده تخصیص داده می‌شوند.حالاتی با مقدار از دست رفته (system-missing) از آنالیز کنار گذاشته می‌شوند. هر مقدار از دست رفته کاربر (user-missing)، در تفکیک متغیرها به‌عنوان موجود در نظر گرفته می‌شود.

## Architecture (ساختار)




*شکل ۲-۱۱: تابع شعاع مدار: نوار ساختار*

نوار ساختار برای تعیین ساختار شبکه مورد استفاده قرار می‌گیرد.این فرایند یک شبکه عصبی با یک لایه"تابع شعاع مدار" پنهان می‌سازد. در حالت کلی تغییر این تنظیمات ضروری نیست.

**تعداد واحدها در لایه پنهان**. سه راه برای انتخاب تعداد واحدهای پنهان وجود دارد.

۱. بهترین تعداد واحد را بین یک بازه محاسبه شده خودکار پیدا کن. فرایند به‌طور خودکار مقادیر حداقل و حداکثر بازه را محاسبه کرده و بهترین تعداد واحد پنهان در این بازه را پیدا می‌کند.

اگر نمونه آزمایش مشخص شده باشد، فرایند از معیار نمونه آزمایش استفاده می‌کند: بهترین تعداد واحد پنهان وقتی است که کوچکترین خطا در داده آزمایش حاصل شود. اگر

نمونه آزمایش مشخص نشود، فرایند از معیار اطلاعات Bayesian (BIC) استفاده می‌کند: بهترین تعداد واحد پنهان وقتی است که کوچکترین BIC براساس داده‌های آموزش حاصل شود.

۲. بهترین تعداد واحدها در یک بازه مشخص را پیدا کن. شما می‌توانید بازه خود را ارائه کنید و فرایند "بهترین" تعداد واحدهای پنهان در آن بازه را پیدا می‌کند. مشابه حالت قبل، بهترین تعداد واحد پنهان بازه با استفاده از معیار داده آزمایش یا BIC مشخص می‌شود.

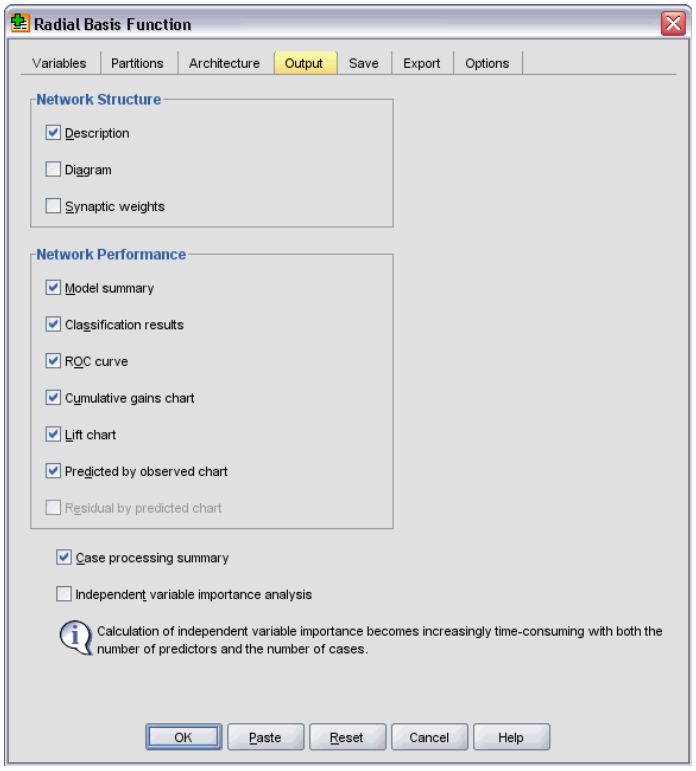
۳. از تعداد مشخصی واحد استفاده کن. شما می‌توانید استفاده از بازه را لغو کرده و مستقیماً تعداد خاصی واحد مشخص کنید.

تابع فعال‌کننده لایه پنهان. تابع فعال‌کننده لایه پنهان تابع شعاع مداری است که واحدهای یک لایه را به مقادیر واحدها در لایه بعد مرتبط می‌کند. برای لایه خروجی تابع فعال‌کننده، تابع همانی است.بنابراین واحدهای خروجی حاصل جمع وزن‌دهی شده ساده واحدهای پنهان هستند.

- تابع شعاع مدار نرمال شده. از تابع فعال‌کننده softmax استفاده می‌کند بنابراین فعال شده تمامی واحدهای پنهان به‌گونه‌ای نرمال می‌شود که حاصل جمع آنها ۱ شود.
- تابع شعاع مدار معمولی. از تابع فعال‌کننده توانی استفاده می‌کند بنابراین فعال شده واحدهای پنهان یک "bump" گوشی به‌صورت تابعی از ورودی‌ها است.

همپوشانی بین واحدهای پنهان. ضریب همپوشانی یک افزایش‌دهنده است که بر عرض تابع شعاع مدار اعمال می‌شود. مقدار محاسبه شده خودکار ضریب همپوشانی ۱+۰,۱d است، که در آن d تعداد واحدهای ورودی (حاصل جمع تعداد دسته‌ها بین تمام ضریب‌ها و تعداد متغیرهای کمکی) است.

## OutPut (خروجی)




*شکل ۲-۱۲: تابع شعاع مدار:نوار خروجی*

**ساختار شبکه**. خلاصه اطلاعات شبکه عصبی را نشان می‌دهد.

- **توصیف**. اطلاعات شبکه عصبی شامل متغیرهای وابسته، تعداد واحدهای ورودی و خروجی، تعداد واحدها و لایه‌های پنهان و توابع فعال کننده را نشان می‌دهد.
- **دیاگرام**. دیاگرام شبکه را به‌صورت یک نمودار غیرقابل تغییر نشان می‌دهد. قابل ذکر است هر چه تعداد متغیر کمکی و ضریب سطوح افزایش یابد، تفسیر دیاگرام دشوارتر می‌شود.
- **وزن‌های سیناپسی**. ضریب‌هایی که برای نشان دادن رابطه بین واحدهای یک لایه و لایه بعد تخمین‌زده شده‌اند نشان می‌دهد. حتی اگر بانک اطلاعات فعال به داده‌های آموزش، آزمایش و جدانگه داشته شده تقسیم شود، وزن‌های سیناپسی براساس داده‌های آموزش تعیین می‌شوند شایان ذکر است که تعداد وزن‌های سناپسی میتواند زیاد باشد و معمولاً از این وزن‌ها برای تفسیر شبکه عصبی استفاده نمی‌شود.

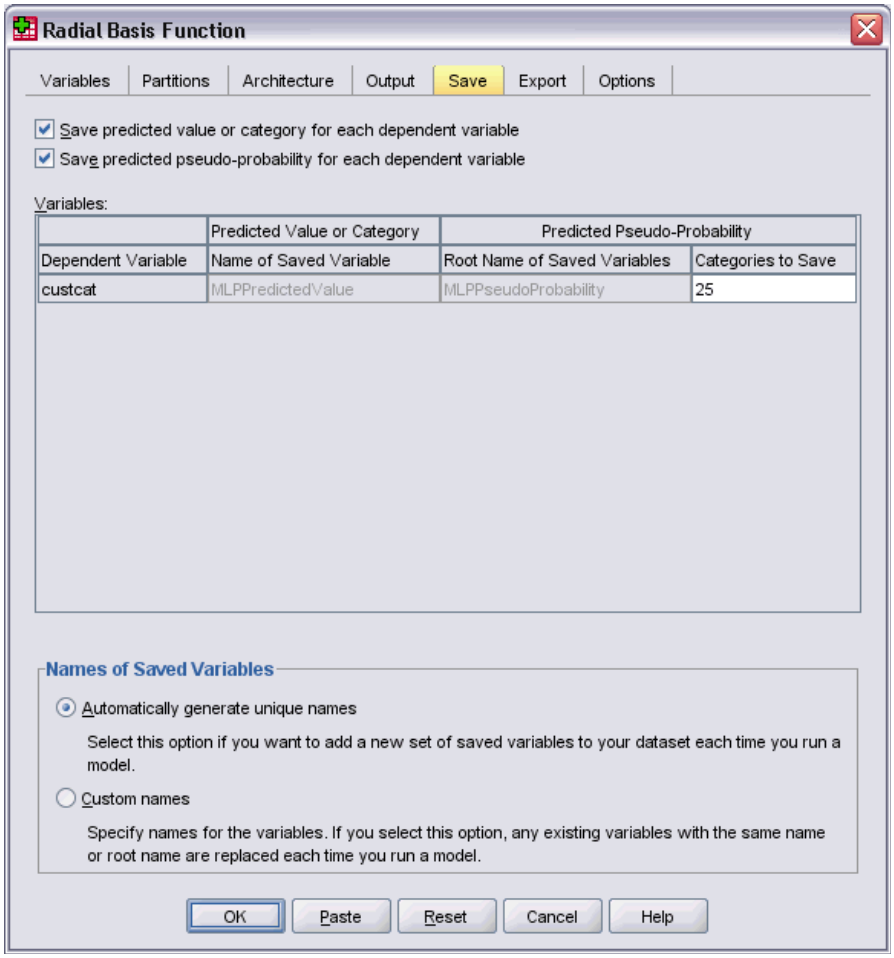
**عملکرد شبکه**. نتایجی را که جهت نشان دادن "خوب بودن" مدل مورد استفاده قرار می‌گیرند نمایش می‌دهد.

**نکته:** نمودارهای این گروه براساس نمونه‌های آموزش و آزمایش یا در صورت عدم وجود نمونه آزمایش تنها بر اساس نمونه آموزش است.

- **خلاصه مدل**. خلاصه‌ای از نتایج شبکه عصبی به تفکیک و شامل همه موارد من جمله خطا، خطای نسبی یا درصد پیش‌بینی نادرست، قوانین توقف جهت متوقف کردن آموزش و زمان آموزش، ارائه می‌دهد.
- از خطای حاصل جمع مربعات استفاده می‌شود. به علاوه، خطای نسبی یا درصد پیش‌بینی نادرست بسته به سطوح اندازه‌گیری متغیر وابسته نمایش داده می‌شود. اگر یکی از متغیرهای وابسته سطح اندازه‌گیری مقیاس‌بندی شده داشته باشد، میانگین خطای نسبی کلی (مربوط به مدل میانگین) نمایش داده می‌شود. اگر تمامی متغیرهای وابسته مطلق باشند، میانگین درصد پیش‌بینی نادرست نمایش داده می‌شود.خطاهای نسبی یا درصدهای پیش‌بینی نادرست برای تک‌تک متغیرهای وابسته نیز مشخص می‌شوند.
- **نتایج طبقه‌بندی**. یک جدول طبقه‌بندی برای هر متغیر وابسته مطلق به‌صورت جزئی و کلی ارائه می‌دهد. هر جدول تعداد حالت‌هایی که برای هر متغیر وابسته مطلق درست یا نادرست طبقه‌بندی شده‌اند، می‌دهد. درصد کل حالت‌هایی که درست طبقه‌بندی شده‌اند نیز گزارش می‌شود.
- **منحنی ROC**. برای هر متغیر وابسته مطلق منحنی ROC نمایش می‌دهد. همچنین جدولی که نمایانگر سطح زیر هر منحنی است ارائه می‌شود. برای یک متغیر وابسته داده شده، نمودار ROC هر دسته با یک منحنی نمایش داده می‌شود. اگر متغیر وابسته ۲ دسته داشته باشد، هر منحنی با دسته‌ی مربوطه به‌صورت حالت مثبت در مقابل دسته‌ی دیگر فرض می‌شود. اگر متغیر وابسته بیش از ۲ دسته داشته باشد، هر منحنی با دسته‌ی مربوطه به‌صورت حالت مثبت در مقابل سایر دسته‌ها فرض می‌شود.
- **نمودار بهره تجمعی**. یک نمودار بهره تجمعی برای هر متغیر وابسته مطلق نمایش داده می‌شود. در اینجا نیز مشابه منحنی‌های ROC ، برای هر دسته‌ی متغیر وابسته یک منحنی ارائه می‌شود.

- **نمودار lift** . یک نمودار lift برای هر متغیر وابسته‌ی مطلق نمایش داده می‌شود. در اینجا نیز مشابه منحنی ROC ، برای هر دسته‌ی متغیر وابسته یک منحنی ارائه می‌شود.
- **نمودار پیش‌بینی شده در مقابل مشاهده شده**. یک نمودار مقادیر پیش‌بینی شده در مقابل مشاهده شده برای هر متغیر وابسته ارائه می‌دهد. برای متغیرهای وابسته مطلق، منحنی جعبه‌ای خوشه‌ای شبه احتمال پیش‌بینی شده، که در آن دسته‌ی پاسخ مشاهده شده متغیر خوشه‌ای است، برای هر پاسخ دسته نمایش داده می‌شود. برای متغیرهای وابسته به مقیاس یک نمودار پراکندگی ارائه می‌شود.
- **منحنی باقی مانده در مقابل پیش‌بینی شده**. یک نمودار مقادیر باقی مانده در مقابل پیش‌بینی شده برای هر متغیر وابسته به مقیاس ارائه می‌دهد. هیچ ساختار مشخصی نباید بین مقادیر باقی مانده و پیش‌بینی شده وجود داشته باشد. این نمودار تنها برای متغیرهای وابسته به مقیاس ترسیم می‌شود.
- **خلاصه پردازش حالت**. جدول خلاصه پردازش حالت که تعداد حالات در برگرفته شده و مستثنی در تحلیل را که به‌صورت کلی و نمونه‌های آموزش، آزمایش و جدا نگه داشته شده خلاصه کرده است، نمایش می‌دهد.
- **تحلیل اهمیت متغیر مستقل**. آنالیز حساسی است که اهمیت هر پیش‌بینی‌کننده در تعیین شبکه عصبی را محاسبه می‌کند. تحلیل ممکن است برپایه نمونه‌های آموزش و آزمایش تلفیق شده یا در صورت عدم وجود نمونه آزمایش تنها روی نمونه آموزش انجام شود. در نهایت یک جدول و یک نمودار که نشان‌دهنده اهمیت و اهمیت نرمال شده هر پیش‌بینی‌کننده است، ارائه می‌شود. شایان ذکر است که تحلیل حساسیت در صورت وجود تعداد زیاد پیش‌بینی‌کننده‌ها یا حالات گران قیمت و زمان‌بر است.

## ذخیره (save)




*شکل ۲-۱۳: تابع شعاع مدار: نوار ذخیره*

نوار ذخیره جهت ذخیره پیش‌بینی‌ها به‌صورت متغیر در بانک اطلاعات مورد استفاده قرار می‌گیرد.

- **مقادیر پیش‌بینی شده یا دسته هر متغیر وابسته را ذخیره کن**. این گزینه مقادیر پیش‌بینی شده برای متغیرهای وابسته به مقیاس و دسته پیش‌بینی شده برای متغیرهای وابسته مطلق را ذخیره می‌کند.

- **شبه احتمال پیش‌بینی شده یا دسته هر متغیر وابسته را ذخیره کن**. این گزینه شبه احتمال پیش‌بینی شده برای متغیر وابسته مطلق را ذخیره می‌کند. یک متغیر جداگانه برای هر n دسته‌ی اول، که n در ستون ( categories to save) تعیین می‌شود، ذخیره خواهد شد.

**نام متغیرهای ذخیره شده**. تولید نام خودکار تضمین می‌کند که شما تمامی کار خود را حفظ کرده‌اید.

نام‌های مرسوم به شما امکان می‌دهد که نتایج را از اجراهای قبلی بدون پاک کردن متغیرهای ذخیره شده در ویرایشگر داده حذف یا جایگرین کنید.

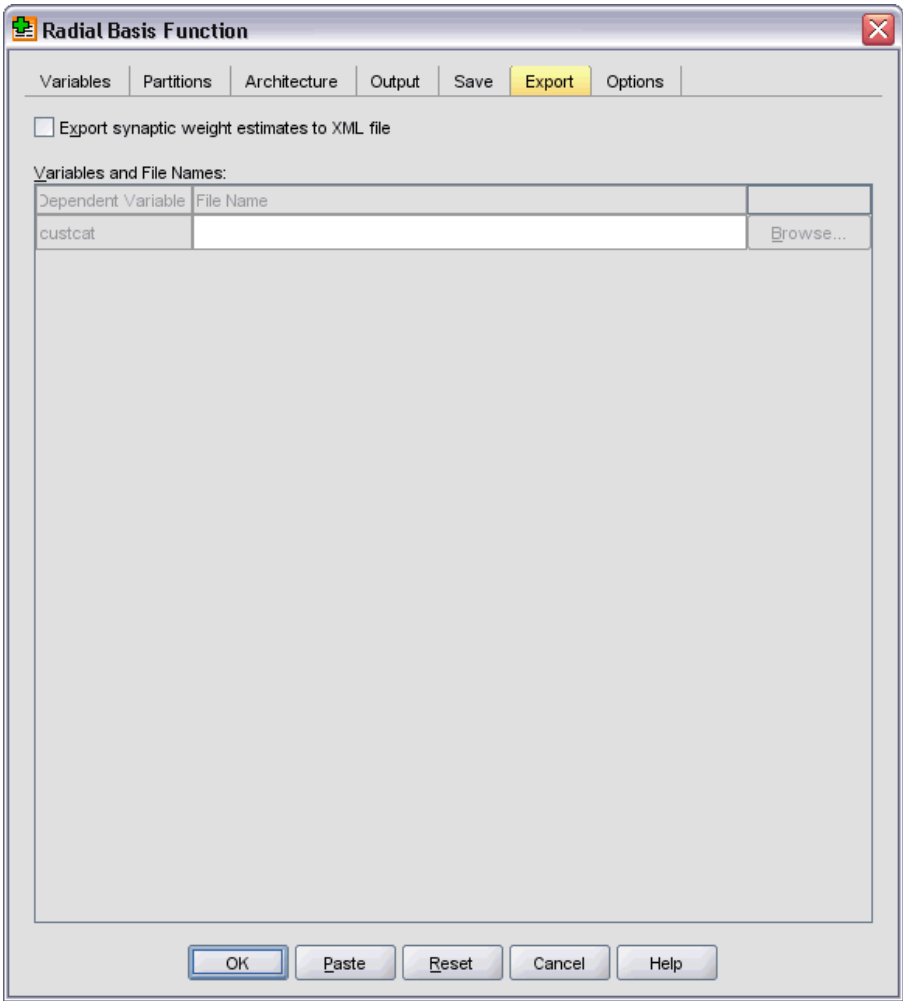
## احتمال‌ها و شبه احتمال‌ها

از آنجا که فرایند تابع شعاع مدار از خطای حاصل جمع مر بعات و تابع فعال‌کننده همانی برای لایه خروجی استفاده می‌کند، نمی‌توان شبه احتمال را به‌عنوان احتمال تفسیر کرد. برنامه این مقادیر شبه احتمال پیش‌بینی شده را نیز حتی اگر کمتر از صفر یا بزرگتر از ۱ بوده و یا حاصل جمع مقادیر یک متغیر وابسته خاص ، ۱ نباشد ، ذخیره می‌کند.

ROC ، بهره تجمعی و نمودار Lift براساس شبه احتمال‌ها ساخته شده‌اند. در حالتی که هر کدام از شبه احتمال‌ها کمتر از صفر یا بزرگتر از ۱ باشند، یا حاصل جمع برای یک متغیر ۱ نباشند، این مقادیر مقیاس‌بندی مجدد می‌شوند تا در بازه ی صفر تا ۱ قرار گرفته و حاصل جمعشان ۱ شود. شبه احتمال با تقسیم شدن به حاصل جمع‌شان مقداردهی مجدد می‌شوند. برای مثال، وقتی برای یک متغیر وابسته سه دسته‌ای مقادیر شبه احتمال پیش‌بینی شده ۰/۵ ، ۰/۶ و ۰/۴ باشد، هر کدام از اینها به ۱/۵ تقسیم می‌شود تا ۰/۳۳ ، ۰/۴ و ۰/۲۷ را بدهد.

اگر هر یک از شبه احتمال‌ها منفی باشد، قدر مطلق کمترین عدد قبل از مقیاس‌بندی مجدد به همه مقادیر اضافه می‌شود. برای مثال اگر شبه احتمال‌ها ۰/۳- ، ۰/۵- و ۱/۳ باشند، ابتدا ۰/۳ به هر مقدار اضافه می‌شود تا ۰ ، ۰/۸ و ۱/۶ به‌دست آید.سپس هر مقدار به حاصل جمع ۲/۴ تقسیم شده تا به ۰ ، ۰/۳۳ و ۰/۶۷ برسیم.

## Export (صدور)




*شکل ۲-۱۴: تابع شعاع مدار:نوار صدور*

نوار Export جهت ذخیره تخمین وزن‌های سیناپسی هر متغیر وابسته برروی یک فایل XML(PMML) استفاده می‌شود.

## Options (گزینه‌ها)


*شکل ۲-۱۵: تابع شعاع مدار: نوار گزینه‌ها*

**مقادیر از دست رفته کاربر.** فاکتورها باید مقادیر قابل دسترسی برای یک حالت داشته باشند تا در تحلیل‌ها وارد شوند. این کنترل‌ها به شما امکان می‌دهد تصمیم بگیرید که مقادیر از دست رفته کاربر در فاکتورها و متغیرهای وابسته مطلق موجود باشند.

# فصل سوم

# مثال‌ها

# بخش اول

## پرسپترون چندلایه

از پرسپترون چند لایه برای ساخت یک مدل پیش‌بینی که در آن به پیش‌بینی یک یا چند متغیر وابسته (هدف) می‌پردازند، استفاده می‌شود.

### استفاده از پرسپترون چندلایه برای محاسبه ریسک حساب‌های اعتباری

یک مامور بخش اعطاء وام که در بانک مشغول به فعالیت می‌باشد، نیازمند آن است تا بتواند ویژگی‌ها و خصوصیات افرادی که ممکن است در باز پرداخت وام غفلت و تأخیر ورزند را شناخته تا با استفاده از آنها میزان ریسک حساب‌های اعتباری آنان را بشناسد.

اطلاعات مربوط به ۸۵۰ نفر از مشتریان سابق و آینده (بالقوه) بانک می‌بایست در بخش bankloan.sav موجود باشد. از ۷۰۰ مورد از این مشتریان برای ساخت یک پرسپترون چند لایه استفاده نمایید و مابقی نمونه‌ها را به‌منظور ارزیابی تحلیل‌های خروجی از سیستم به کنار بگذارید. سپس به‌عنوان نمونه از مدل جهت دسته‌بندی ۱۵۰ نفر از مشتریان آتی بانک، برای تشخیص مشتریان خوب یابد استفاده نموده و میزان ریسک حساب‌های اعتباری آنان را محاسبه نمایید.

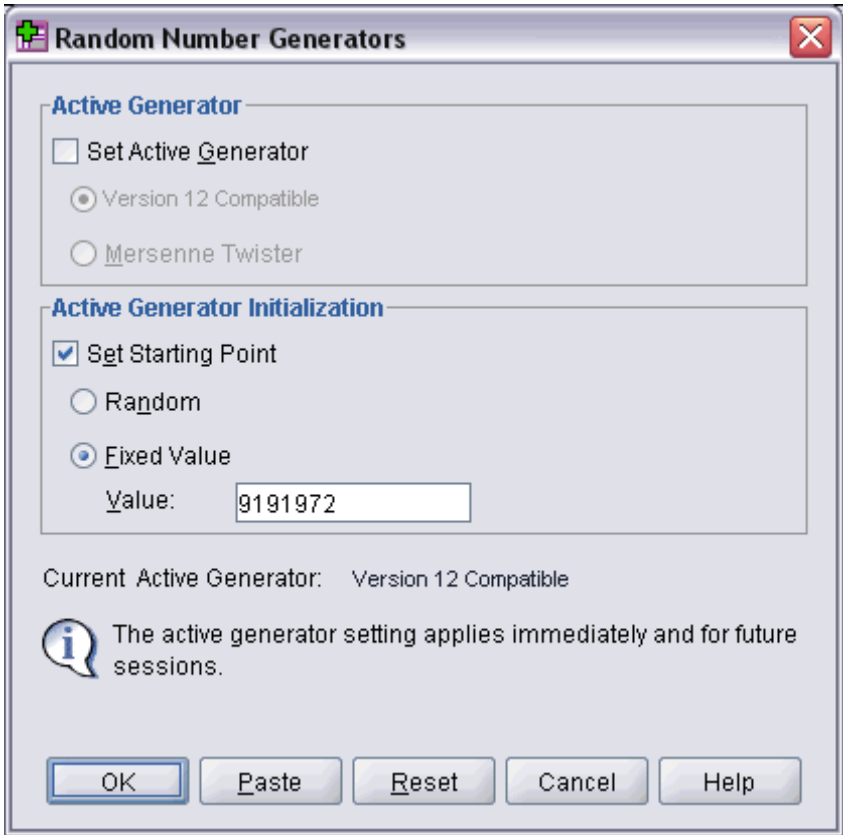
به علاوه، مسئول اعطاء وام، پیش داده‌ها را با استفاده از رگرسیون مورد تحلیل قرار داده است و می‌بداند که چگونه می‌توان از پرسپترون‌های چند لایه‌ای مانند یک ابزار دسته‌بندی‌کننده نیز استفاده نمود.

## آماده‌سازی داده‌ها جهت انجام تحلیل‌ها

ایجاد دسته‌بندی‌های تصادفی این اجازه را به شما می‌دهد که شرایط انجام تحلیل‌ها را به دقت و درستی فراهم آورید.

- برای ایجاد دسته‌بندی‌های تصادفی، از منو، انتخاب نمایید:

Transform→Random number generators…



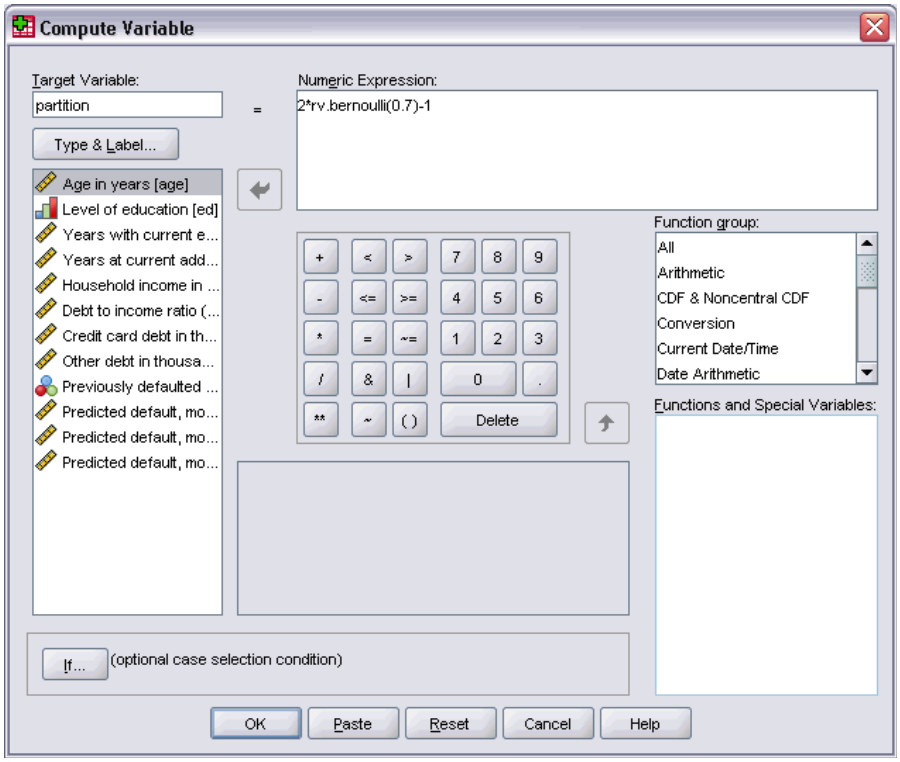

شکل ۳-۱

- گزینه set starting point را انتخاب نمایید.
- گزینه Fixed value را انتخاب نموده و عدد ۹۱۹۱۹۷۲ را برای مقدار خواسته شده تایپ نمایید.
- گزینه ok را کلیک کنید.

در تحلیل‌های رگرسیونی که از پیش انجام گرفته شده است، مشخص گردید که در حدود ۷۰٪ از مشتریان سابق، نمونه‌های آموزشی قلمداد شده‌اند و ۳۰٪ از آنها به دسته نمونه‌های جدا نگه داشته شده (Holdout) تعلق پیدا کرده‌اند. استفاده از یک متغیر تفکیک‌کننده جهت دوباره‌سازی نمونه‌های استفاده شده در آن تحلیل‌ها، موردنیاز خواهد بود.

- برای ساخت متغیر تفکیک‌کننده از قسمت منو، انتخاب کنید:

Transform→Compute variable




شکل ۳-۲

- کلمه Partition را در قسمت Target variable تایپ نمایید.
- عبارت 2*rv.bernoulli(0.7)-1 را در قسمت Numeric Expression تایپ نمایید.

این فرایند در نتیجه باعث خواهد شد، مجموعه مقادیر تفکیکی به‌صورت تصادفی از یک تابع توزیع برنولی با پارامتر احتمال ۰/۷ دسته‌بندی گردند، که در نتیجه آن مقادیر ۱ و ۱- را به جای ۱ و ۰ به خود خواهد گرفت. بدین‌وسیله تمامی مواردی که مقادیر مثبت به خود می‌گیرند را به‌عنوان نمونه‌های training (آموزشی) و مواردی که دارای مقادیر منفی می‌شوند را به‌عنوان نمونه‌های جدانگه داشته شده و مواردی که مقدار صفر را به خود می‌گیرند را به‌عنوان نمونه‌های Testing (آزمایش) در نظر خواهیم گرفت. در حال حاضر قصد شاخص نمودن نمونه‌های testing (آزمایش) را نداریم.

- برروی گزینه ok در پنجره Compute variable کلیک کنید.

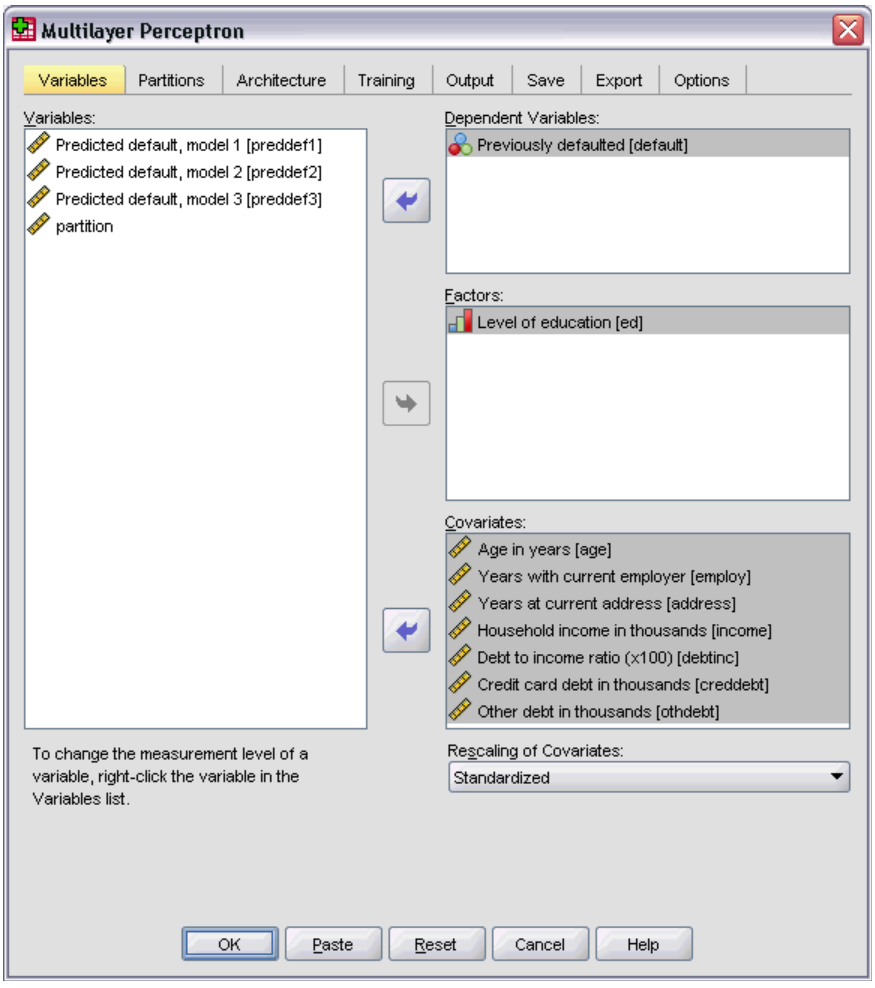
در حدود ۷۰٪ از مشتریانی که پیش از این وام دریافت کرده‌اند مقدار ۱ را توسط متغیر تفکیکی به خود اختصاص داده‌اند. از این مشتریان برای ساخت مدل استفاده خواهد شد. باقی

مشتریانی که پیش از این وام دریافت نموده‌اند مقدار ۱- را دریافت خواهند کرد و از آنها جهت ارزیابی نتایج مدل استفاده خواهد شد.

## شروع تحلیل‌ها

برای شروع تحلیل‌ها در یک پرسپترون چند لایه می‌بایست، از منو انتخاب نمایید:

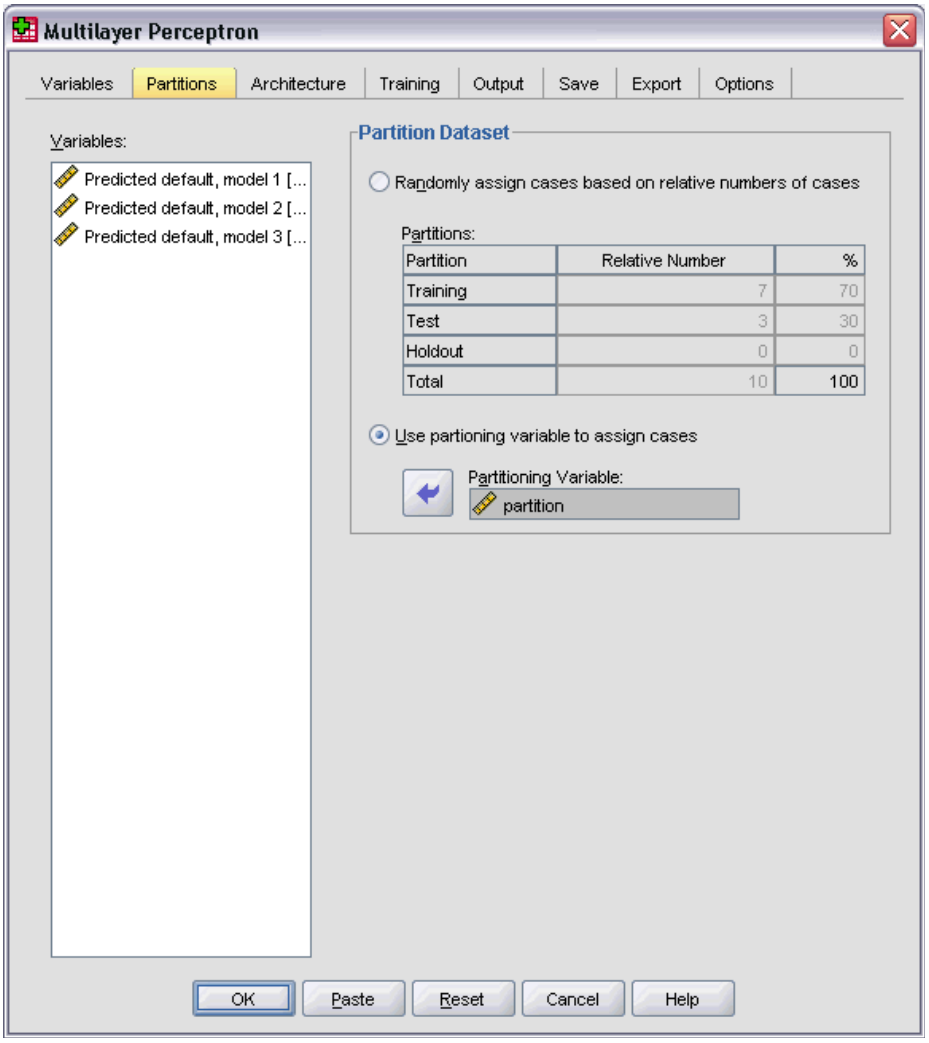
Analyze→Neural Networks→Multilayer Perceptron




*شکل ۳-۴*

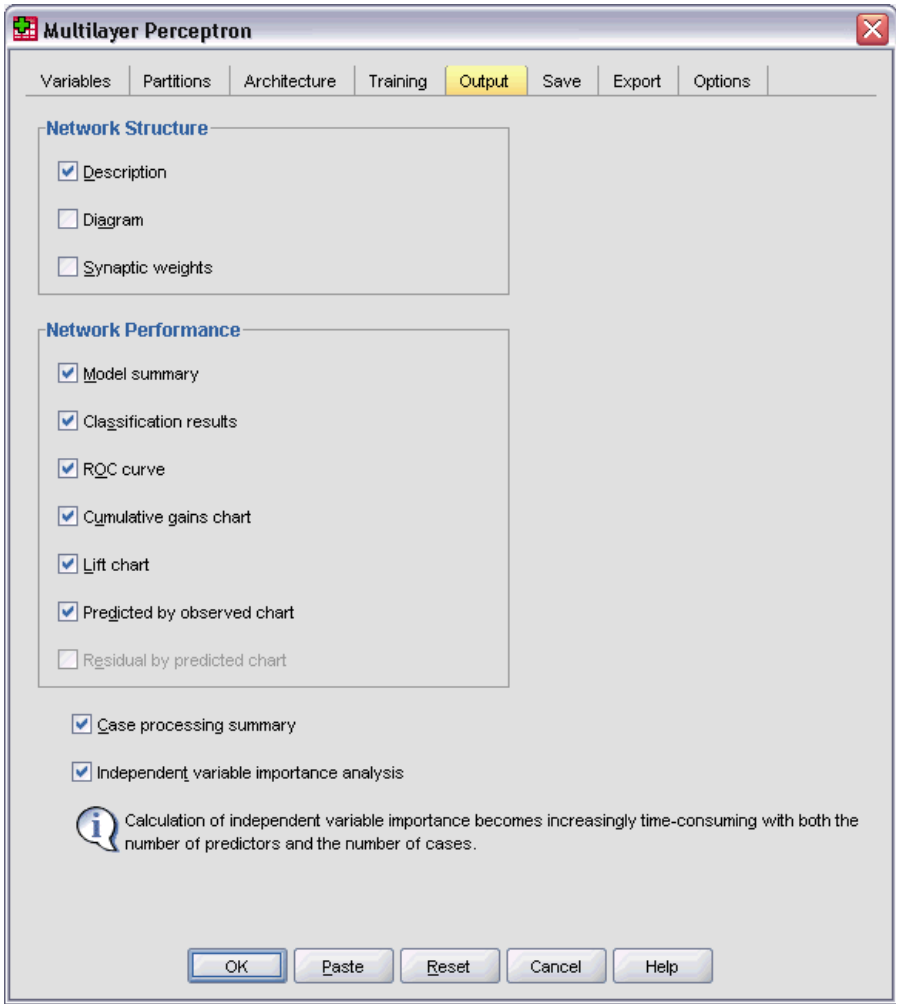
- Previously defaulted(default) را به‌عنوان متغیر وابسته انتخاب نمایید.
- Level of education(ed) را به‌عنوان Factor انتخاب نمایید.

- Age in years(age) تا other debt in thousands (othdebt) را در قسمت متغیر کمکی انتخاب نمایید.




شکل ۳-۵

- گزینه use partitioning variable to را انتخاب کنید.
- عبارت Partition را به‌عنوان Partitioning variable انتخاب نمایید.
- برروی گزینه output کلیک کنید.




شکل ۳-۶

- علامت تیک را در قسمت Network Structure از کنار گزینه Diagram بردارید.
- گزینه‌های ROC curve ، Cummulative gains chart ، Lift chart ، Predicted by observed chart را در گروه Network Performance انتخاب نمایید. همان‌طورکه مشاهده می‌نمایید گزینه Residual by predicted chart به‌صورت غیرفعال می‌باشد، این مسئله بدین دلیل است که متغیر وابسته هنوز مقیاس‌بندی نگردیده است.
- گزینه independent variable importance analysis را انتخاب نمایید.
- برروی Ok کلیلک نمایید.

## خلاصه فرایند انجام شده

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 499 | 71.3% |
| | Holdout | 201 | 28.7% |
| Valid | | 700 | 100.0% |
| Excluded | | 150 | |
| Total | | 850 | |

شکل ۳-۷

خلاصه فرایند انجام شده نشان می‌دهد که ۴۹۹ مورد از مشتریان در گروه نمونه‌های Training و ۲۰۱ مورد در گروه نمونه‌های Holdout قرار گرفته‌اند. ۱۵۰ موردی که از تحلیل‌ها خارج شده‌اند، مشتریان بالقوه بانک در آینده می‌باشد.

## اطلاعات شبکه

| | | | |
|---|---|---|---|
| Input Layer | Factors | 1 | Level of education |
| | Covariates | 1 | Age in years |
| | | 2 | Years with current employer |
| | | 3 | Years at current address |
| | | 4 | Household income in ... |
| | | 5 | Debt to income ratio (x100) |
| | | 6 | Credit card debt in thousands |
| | | 7 | Other debt in thousands |
| | Number of Units[a] | | 12 |
| | Rescaling Method for Covariates | | Standardized |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1[a] | | 4 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | Previously defaulted |
| | Number of Units | | 2 |
| | Activation Function | | Softmax |
| | Error Function | | Cross-entropy |

a.Excluding the bias unit

شکل ۳-۸

جدول Network Information ، اطلاعات مربوط به شبکه‌های عصبی را نشان داده و برای حصول اطمینان از اینکه موارد اختصاص یافته صحیح می‌باشند، کاربرد دارد. اکنون به موارد زیر توجه نمایید:

- تعداد واحدهای قرار گرفته در لایه ورودی برابر با تعداد متغیر کمکی به علاوه، تعداد کل سطوح عامل (Factor levels) می‌باشد. واحدی مجزا برای دسته، سطح تحصیلات (Level of education)

به وجود آمده و هیچ یک از دسته‌بندی‌ها، بدین‌صورت‌که در بسیاری از مدل‌های دیگر زائد و اضافی در نظر گرفته می‌شوند، در این جا زائد در نظر گرفته نمی‌شود.

- به همین ترتیب، واحد خروجی مجزایی نیز برای هریک از گروه‌های که در گذشته بدحسابی داشته‌اند ساخته شده است (Previously defaulted) که این میزان در کل ۲ واحد در لایه خروجی است.
- ساختار گزینش اتوماتیک، ۴ واحد را در لایه پنهان انتخاب کرده است.
- باقی اطلاعات مربوط به شبکه به‌صورت پیش‌فرض در فرایند در نظر گرفته شده است.

## خلاصه مدل

| | | |
|---|---|---|
| Training | Cross Entropy Error | 156.606 |
| | Percent Incorrect Predictions | 15.6% |
| | Stopping Rule Used | Maximum number of epochs (100) exceeded |
| | Training Time | 00:00:00.081 |
| Holdout | Percent Incorrect Predictions | 25.4% |

Dependent Variable: Previously defaulted

*شکل ۳-۹*

در بخش Model summary اطلاعات مربوط به نتایج آموزش انجام شده و استفاده از شبکه نهایی در مورد نمونه‌های جدانگه داشته شده (Holdout) را می‌توانید مشاهده نمایید.

- خطای آنتروپی (Cross entropy error) به دلیل آن که در لایه خروجی از تابع فعال‌کننده softmax استفاده شده است، در جدول آورده شده است. شبکه در طی فرایند خود سعی دارد میزان این تابع خطا را به کمترین مقدار خود برساند.
- میزان پیش‌بینی‌های اشتباه که به‌صورت درصد نشان داده می‌شوند، از جدول طبقه‌بندی (classification table)، که در ادامه بیشتر پیرامون آن بحث خواهیم کرد، گرفته شده است.
- الگوریتم محاسبه به دلیل آن که به میزان حداکثری از تعداد دفعات شروع دوره‌ها رسیده‌ایم، متوقف شده است. در حالت ایده‌آل این الگوریتم می‌بایست به دلیل همگرا شدن خطاها متوقف گردد. این نکته، تردیدها را در مورد این که در طول آموزش در مواردی

اشتباه رخ داده است و یا این که در ادامه و در زمانی که بررسی‌های بیشتر برروی خروجی‌ها صورت می‌گیرد، این اشتباهات رخ خواهد داد را افزایش می‌دهد.

## طبقه‌بندی

| Sample | Observed | Predicted No | Predicted Yes | Percent Correct |
|---|---|---|---|---|
| Training | No | 347 | 28 | 92.5% |
| | Yes | 50 | 74 | 59.7% |
| | Overall Percent | 79.6% | 20.4% | 84.4% |
| Holdout | No | 123 | 19 | 86.6% |
| | Yes | 32 | 27 | 45.8% |
| | Overall Percent | 77.1% | 22.9% | 74.6% |

Dependent Variable: Previously defaulted

*شکل ۳-۱۰*

جدول Classification نتایج خاصی که حاصل از استفاده از شبکه می‌باشد را نشان می‌دهد. در هر مورد، درصورتی‌که شبه احتمال پیش‌بینی، بیشتر از ۰/۵ باشد، عکس‌العمل سیستم مثبت خواهد بود.

- سلول‌های قرار گرفته برروی diagonal of the cross –classification در هر مورد، پیش‌بینی‌های صحیح می‌باشد.
- سلول‌هایی که برروی diagonal of the cross – classification قرار نگرفته‌اند، پیش‌بینی‌های غیرصحیح می‌باشند.

در بین مواردی که از آنها برای ساخت مدل استفاده شده است، ۷۴ مورد از ۱۲۴ موردی که در گذشته بدحسابی داشته‌اند، به درستی طبقه‌بندی شده‌اند، و تعداد ۳۴۷ مورد از ۳۷۵ موردی که بدحسابی نداشته‌اند نیز به درستی طبقه‌بندی شده‌اند. به‌صورت کلی ۸۴/۴٪ از موارد آموزش به درستی طبقه‌بندی شده و میزان باقی مانده یعنی ۱۵/۶٪ موارد اشتباه در جدول خلاصه مدل نشان داده شده است. عملکرد مدلی مناسب‌تر ارزیابی می‌گردد که درصد بیشتری از موارد صحیح را نشان داده و مشخص نماید.

دسته‌بندی‌هایی که در آنها از مواردی جهت ساخت مدل استفاده می‌گردد، می‌بایست به میزان زیادی خوشبینانه در نظر گرفته شوند، به این معنا که می‌بایست نسبت طبقه‌بندی آنها

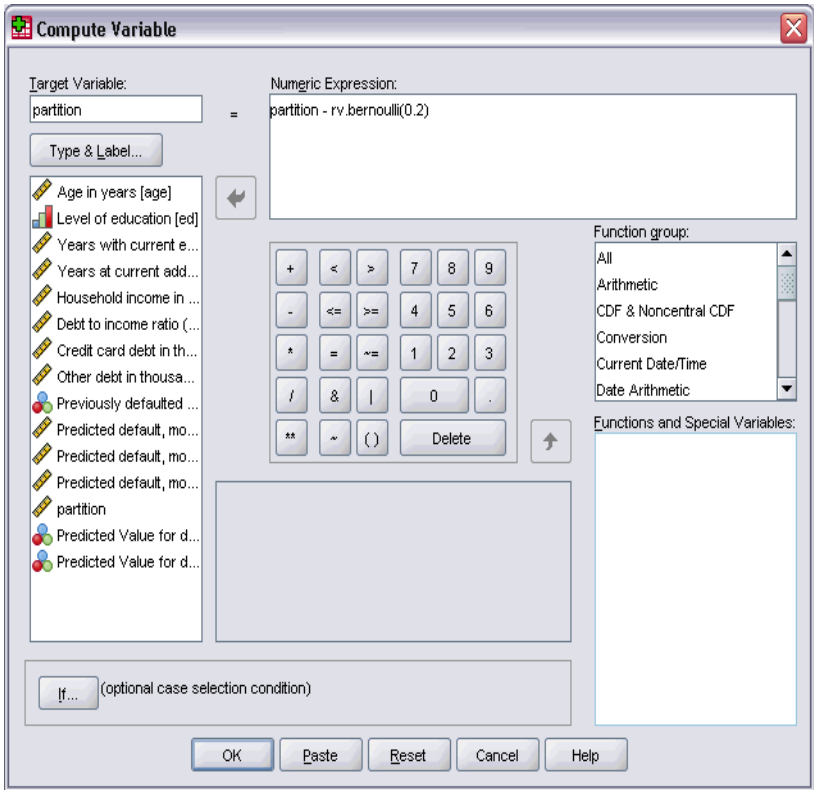
زیاد باشد. نمونه‌های جدا نگه داشته شده، به ارزیابی مدل کمک می‌کنند. در این جا ۷۴/۶٪ از موارد به درستی طبقه‌بندی گردیده‌اند. این مسئله بیانگر آن است که در حالت کلی، مدل شما در واقعیت۳ بار از ۴ بار صحیح عمل می‌نماید.

## تصحیح نمودن آموزش اضافی

بار دیگر به نتایج تحلیل‌های رگرسیونی که پیش از این ارائه گردیده است، توجه نمایید. مأمور اعطای وام عنوان می‌دارد که نمونه‌های آموزش و جدا نگه داشته شده به درستی، درصد مشابهی از موارد را پیش‌بینی نموده‌اند که این رقم ۸۰٪ می‌باشد. شبکه عصبی در مورد نمونه‌های آموزش درصد بالاتری از پیش‌بینی‌های درست را به همراه داشته اما در مورد نمونه‌های جدا نگه داشته شده نتیجه‌ها مناسب نبوده و درصد اعلام‌های درست به میزان قابل توجهی کم است. (۴۵/۸٪ موارد صحیح در نمونه‌های جدا نگه داشته شده و در ۵۹/۷٪ نمونه‌های آموزشی) این نکته به همراه شرط توقف گزارش شده در بخش خلاصه مدل می‌تواند شما را به این که آموزش اضافی اتفاق افتاده است، مشکوک نماید. اما در واقع این آموزش اضافی حاصل حذف نمودن الگوهای زائدی است که در طی فرایند آموزش و با واریانس‌های تصادفی اتفاق می‌افتد.

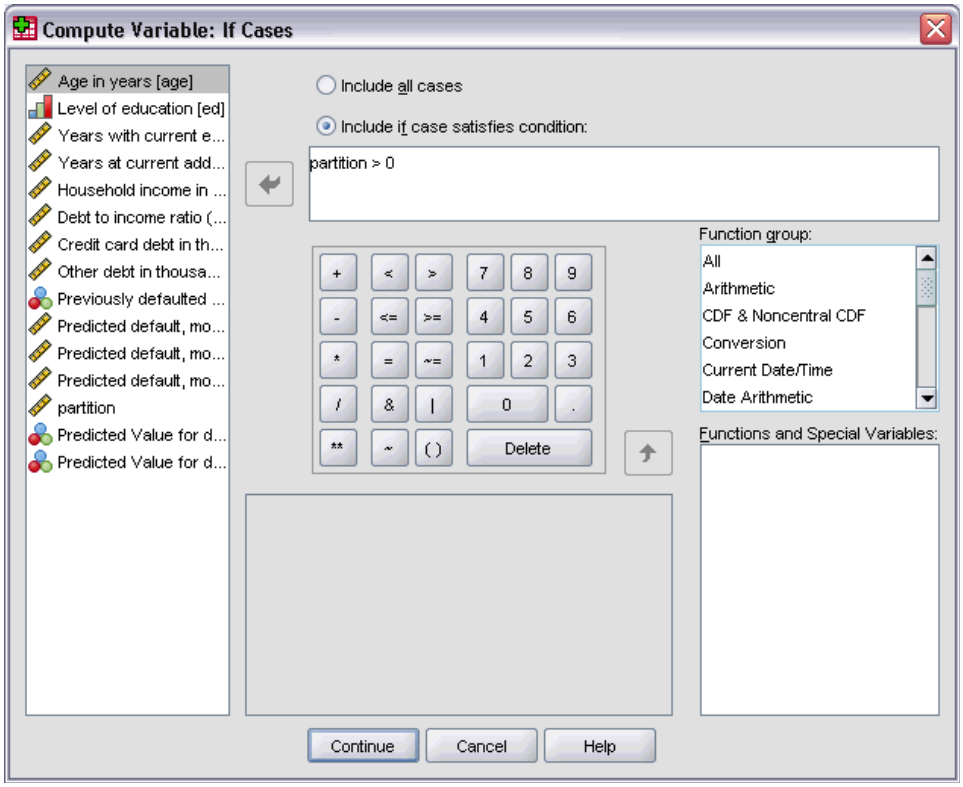
خوشبختانه، راه‌حل این مشکل بسیار ساده است. یک نمونه آزمایشی را که کمک می‌نماید شبکه "on track" باقی بماند را مشخص نمایید. اگر چه که رگرسیون منطقی هیچ جنبه‌ای از نمونه آزمایشی را به همراه ندارد اما متغیر تفکیکی را که پیش از این ساخته‌ایم می‌تواند به دقت نمونه‌های جدا نگه داشته شده و آموزش را که در تحلیل‌های رگرسیونی استفاده شده‌اند بازسازی نمایند. بنابراین بخشی از نمونه آموزشی را گرفته و از آن به‌عنوان یک نمونه آزمایشی استفاده می‌نماییم.

## ایجاد نمونه آموزشی




*شکل ۳-۱۱*

- صفحه Compute variable را بار دیگر باز نمایید.
- در بخش Numeric expression عبارت Partition – rv.bernoulli(0.2) را تایپ نمایید.
- برروی گزینه If کلیک نمایید.




شکل ۳-۱۲

- عبارت if case satisfies condition را انتخاب و علامت‌گذاری نمایید.
- عبارت Partition > 0 را تایپ کنید.
- برروی Continue کلیک کنید.
- برروی گزینه Ok در داخل صفحه Compute Variable کلیک نمایید.

این فرایند، مقادیر نسبت داده شده به Partition (متغیر تفکیکی) را دوباره مقداردهی می‌کند، این کار در مورد مقادیر بزرگتر از صفر انجام می‌گیرد، بدین‌ترتیب در حدود ۲۰٪ از مقادیر صفر شده و ۸۰٪ آنها ۱ باقی می‌مانند. به‌طورکلی، در حدود ۵۶٪ = (۰/۸ × ۰/۷) × ۱۰۰ از مشتریانی که پیش از این وام دریافت نموده‌اند در نمونه آموزشی و ۱۴٪ آنها در نمونه آزمایشی حضور خواهند داشت.

مشتریانی که از ابتدا در گروه نمونه‌های جدا نگه داشته شده، قرار گرفته‌اند، در همان گروه باقی می‌مانند.

## آغاز نمودن تحلیل‌ها

- صفحه Multilayer Perceptron را بار دیگر باز نمایید و برروی گزینه Save کلیک نمایید.
- عبارت save predicted pseudo – probability for each dependent را انتخاب نمایید.
- برروی گزینه Ok کلیک کنید.

## خلاصه‌ای از فرایند انجام شده

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 398 | 56.9% |
| | Testing | 101 | 14.4% |
| | Holdout | 201 | 28.7% |
| Valid | | 700 | 100.0% |
| Excluded | | 150 | |
| Total | | 850 | |

شکل ۳-۱۶

از میان ۴۹۹ نمونه‌ای که در اصل برای قرار گرفتن در نمونه‌های آموزشی مشخص گردیده‌اند، ۱۰۱ مورد برای قرار گرفتن در نمونه آزمایشی دوباره انتخاب می‌شوند.

## اطلاعات شبکه

| | | | |
|---|---|---|---|
| Input Layer | Factors | 1 | Level of education |
| | Covariates | 1 | Age in years |
| | | 2 | Years with current employer |
| | | 3 | Years at current address |
| | | 4 | Household income in ... |
| | | 5 | Debt to income ratio (x100) |
| | | 6 | Credit card debt in thousands |
| | | 7 | Other debt in thousands |
| | Number of Units[a] | | 12 |
| | Rescaling Method for Covariates | | Standardized |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1[a] | | 7 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | Previously defaulted |
| | Number of Units | | 2 |
| | Activation Function | | Softmax |
| | Error Function | | Cross-entropy |

a.Excluding the bias unit

*شکل ۳-۱۷*

تنها تغییری که در جدول اطلاعات شبکه مشاهده می‌گردد، این است که ساختار انتخاب اتوماتیک، ۷ واحد را در لایه پنهان انتخاب نموده است.

## خلاصه مدل

| | | |
|---|---|---|
| Training | Cross Entropy Error | 159.870 |
| | Percent Incorrect Predictions | 20.1% |
| | Stopping Rule Used | 1 consecutive step (s) with no decrease in error[a] |
| | Training Time | 00:00:01.013 |
| Testing | Cross Entropy Error | 40.068 |
| | Percent Incorrect Predictions | 17.8% |
| Holdout | Percent Incorrect Predictions | 20.4% |

Dependent Variable: Previously defaulted

a.Error computations are based on the testing sample.

*شکل ۳-۱۸*

خلاصه مدل، نشان‌دهنده علایم مثبتی است.

- درصد پیش‌بینی‌های نادرست تقریباً برابر با نمونه‌های آموزش و آزمایش و جدا نگه داشته شده می‌باشند.
- الگوریتم محاسبه به دلیل آن که پس از گذر یک مرحله در آن میزان خطا، کاهش نیافته متوقف شده است.

همچنین این جدول نشان می‌دهد که مدل اصلی ممکن است دارای آموزشی اضافی بوده و این مشکل با اضافه نمودن یک نمونه آزمایشی حل شده است. البته، اندازه نمونه‌ها، نسبتاً کوچک بوده و ما نباید در مورد نوسان‌های کم دامنه و چند درصدی چندان نگران باشیم.

## طبقه‌بندی

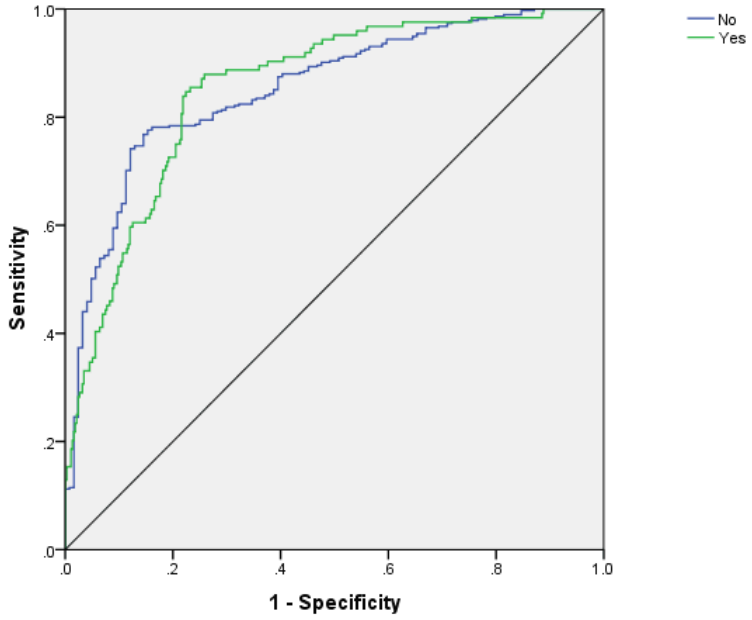
| | | Predicted | | |
|---|---|---|---|---|
| Sample | Observed | No | Yes | Percent Correct |
| Training | No | 263 | 34 | 88.6% |
| | Yes | 46 | 55 | 54.5% |
| | Overall Percent | 77.6% | 22.4% | 79.9% |
| Testing | No | 73 | 5 | 93.6% |
| | Yes | 13 | 10 | 43.5% |
| | Overall Percent | 85.1% | 14.9% | 82.2% |
| Holdout | No | 124 | 18 | 87.3% |
| | Yes | 23 | 36 | 61.0% |
| | Overall Percent | 73.1% | 26.9% | 79.6% |

Dependent Variable: Previously defaulted

*شکل ۳-۱۹*

جدول طبقه‌بندی نشان می‌دهد که با قراردادن عدد ۰٫۵ به‌عنوان شبه احتمال و شرط پایان طبقه‌بندی، شبکه به شکل قابل ملاحظه‌ای در پیش‌بینی خوش حسابان نسبت به بد حسابان بهتر عمل می‌نماید. متأسفانه قرار دادن تنها یک شرط خاتمه طبقه‌بندی به شما دید محدودی از توانایی شبکه در پیش‌بینی را ارائه می‌دهد، بنابراین اساساً چندان برای مقایسه شبکه‌های محاسباتی مناسب نمی‌باشد و می‌توان به جای آن از منحنی ROC استفاده نمود.

## منحنی ROC




شکل ۳-۲۰

منحنی ROC به ارائه تصویری گویا از میزان حساسیت و شاخص نمایی برای تمامی مقادیر ممکن محدودیت توقف، تنها در یک نمودار و به‌طور همزمان می‌پردازد که بسیار واضح‌تر و توانمندتر از مجموعه‌ای از جداول می‌تواند عمل نماید.

نموداری که در این جا نشان داده شده است. دو منحنی را نمایش می‌دهد، یکی از آنها گروه NO و دیگری گروه YES را نمایندگی می‌کنند. از آنجایی که تنها دو دسته‌بندی وجود دارد، منحنی‌ها نسبت به خط ۴۵ درجه (که در نمودار نشان داده نشده است) از گوشه بالایی سمت چپ نمودار تا قسمت پایینی قسمت راست متقارن می‌باشند.

توجه داشته باشید که این نمودار براساس ترکیبی از نمونه‌های آموزشی و آزمایشی استوار است. برای رسم یک نمودار ROC در مورد نمونه‌های جدا نگهداری شده، فایل را توسط متغیر تفکیکی به دو نیم تقسیم نموده و فرایند ترسیم منحنی ROC را با توجه به میزان شبه احتمال مشخص و ذخیره شده، راه‌اندازی می‌نمایید.

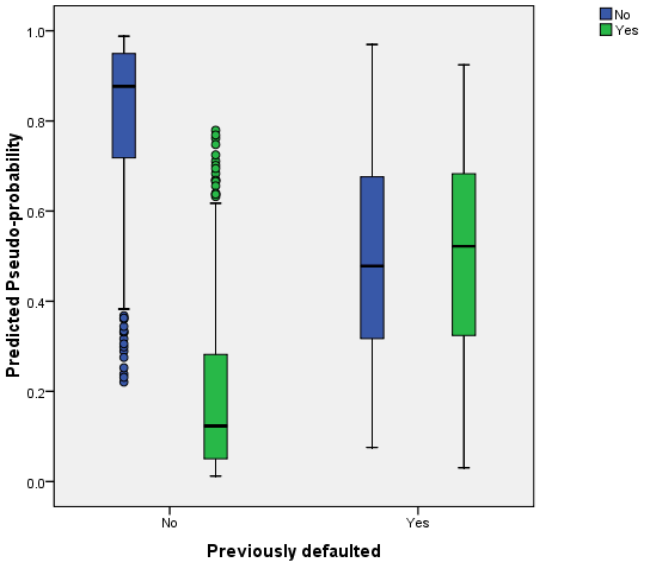
| | | Area |
|---|---|---|
| Previously defaulted | No | .853 |
| | Yes | .853 |

شکل ۳-۲۱

فضای زیر منحنی، خلاصه و چکیده‌ای عددی از منحنی ROC، است و عددهایی که در جدول آمده است، نشان‌دهنده این نکته هستند که در هر دسته، احتمالی که به‌عنوان شبه احتمال برای حضور در دسته مشخصی پیش‌بینی شده است، در مواردی که به‌صورت تصادفی در آن دسته انتخاب شده‌اند نسبت به حالتی که به‌صورت تصادفی اما نه در آن دسته مشخص انتخاب شده‌اند، بیشتر است. برای مثال، در مورد یک شخص بدحساب و یک شخص خوش که به‌صورت تصادفی انتخاب شده‌اند، ۰/۸۵۳ احتمال وجود دارد که مدل پیشگو، شبه احتمال بدحسابی را بیشتر از خوش‌حساب‌ها پیش‌بینی نماید.

از آنجایی که فضای زیر منحنی یک خلاصه آماری مناسب و مفید جهت نمایش صحت عملکرد شبکه ارائه می‌دهد، شما می‌بایست معیار مناسب و خاصی را برای طبقه‌بندی مشتریان انتخاب نمایید. نمودار پیش‌بینی براساس مشاهده (Predicted-by-Observed Chart) می‌تواند نقطه آغاز مناسبی برای این امر باشد.

## نمودار پیش‌بینی براساس مشاهده (Predicted-by-Observed Chart)




شکل ۳-۲۲

نمودار پیش‌بینی براساس مشاهده برای متغیرهای وابسته در هر دسته، متشکل از نمودارهای میله‌ای تجمعی مربوط به شبه احتمالات پیش‌بینی شده است که برای مجموعه‌ای مرکب از نمونه‌های آموزشی و آزمایشی به‌دست آمده‌اند.

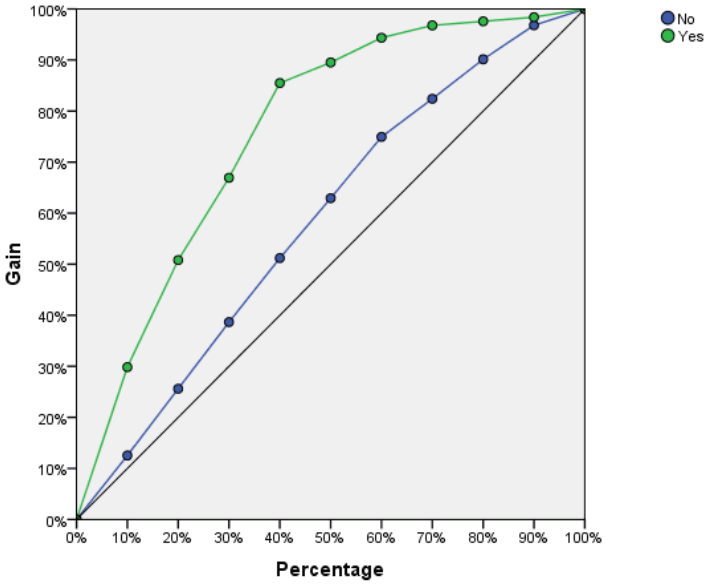
- نمودار میله‌ای که در منتهی‌الیه سمت چپ قرار دارد، نشان‌دهنده میزان شبه احتمال پیش‌بینی شده دسته NO ، برای مواردی است که در دسته NO قرار گرفته‌اند. قسمتی از نمودار میله‌ای که در بالای عدد ۰/۵ (که برروی محور y مشخص شده است) قرار گرفته نشان‌دهنده پیش‌بینی‌های صحیحی است که در جدول طبقه‌بندی آورده شده بود. شبکه در پیش‌بینی مواردی که در دسته NO قرار می‌گیرند و شرط توقف ۰/۵ را برای آنان لحاظ نموده‌ایم، بسیار خوب عمل می‌نماید. بنابراین تنها بخش کوچکی از قسمت انتهایی نمودار نشان‌دهنده تعداد کمی از مواردی است که اشتباه دسته‌بندی شده‌اند.
- نمودار میله‌ای بعدی که با حرکت به سمت راست به آن می‌رسیم، نشان‌دهنده شبه احتمال پیش‌بینی شده دسته YES، برای مواردی است که در دسته NO قرار می‌گیرد. از آنجایی که در این مثال تنها دو دسته‌بندی برای متغیر هدف وجود دارد، دو نمودار میله‌ای اول نسبت به خط افقی ۰/۵ متقارن می‌باشد.
- سومین نمودار میله‌ای نشان‌دهنده شبه احتمال پیش‌بینی شده دسته NO، برای مواردی است که در دسته YES قرار گرفته‌اند. این نمودار و آخرین نمودار میله‌ای نسبت به خط افقی ۰/۵ ، متقارن می‌باشند.
- آخرین نمودار میله‌ای نشان‌دهنده، شبه احتمال پیش‌بینی شده دسته YES، برای مواردی است که در دسته YES قرار گرفته‌اند. قسمتی از این نمودار میله‌ای که در بالای عدد ۰/۵ برروی محور Yها قرار دارند نشان‌دهنده مواردی است که به درستی پیش‌بینی شده‌اند و بخش زیرین این عدد در نمودار نماینده مواردی است که به‌صورت نادرست مورد پیش‌بینی قرار گرفته‌اند. همان‌طورکه از جدول طبقه‌بندی به خاطر دارید، شبکه کمی بیشتر از نیمی از کل موارد را در دسته YES، با شرط توقف ۰/۵ به درستی پیش‌بینی می‌نماید. بنابراین بخش قابل توجهی از موارد به شیوه نادرستی دسته‌بندی شده‌اند.

به نمودار توجه نمایید، به نظر می‌رسد که با کم کردن میزان شرط توقف برای دسته‌بندی یک مورد در دسته YES، از عدد ۰/۵ تا حدود تقریبی ۰/۳ (این مقدار به‌صورت تقریبی برابر است با مقدار بالای دومین نمودار و کف چهارمین نمودار) می‌توانید شانس شناسایی

مشتریانی که در آینده ممکن است بدحساب باشند را افزایش داده و این کار را بدون از دست دادن پتانسیل داشتن مشتریان خوب و خوش حساب انجام دهید.

با حرکت از عدد ۰/۵ به سمت ۰/۳ در طول دومین نمودار میله‌ای تعداد نسبتاً کمی از مشتریان خوش‌حساب به شکل نادرست دسته‌بندی شده و به‌عنوان مشتریان بدحساب شناخته می‌شوند، این درحالی است که در نمودار میله‌ای چهارم این جابه‌جایی سبب دسته‌بندی مجدد بسیاری از مشتریان بدحساب می‌شود که در حالت قبل به شکل نادرست، به‌عنوان مشتریان خوش حساب دسته‌بندی شده بودند.

## Cumulative Gains and lift charts




*شکل ۳-۲۳*

این نمودار نشان می‌دهد که با هدف قرار دادن درصد مشخصی از کل موارد در دسترس، چه درصدی از تعداد موارد را در هر دسته شامل می‌شود. برای مثال، اولین نقطه برروی منحنی دسته YES دارای مختصاتی به‌صورت (۳۰٪ و ۱۰٪ ) می‌باشد که این مختصات بدین معناست که چنانچه شما یک مجموعه‌ای از داده‌ها را به‌وسیله شبکه ثبت نموده و تمامی موارد در دسترستان را با شبه احتمال پیش‌بینی شده YES، منظم نمایید، می‌توانید انتظار داشته باشید

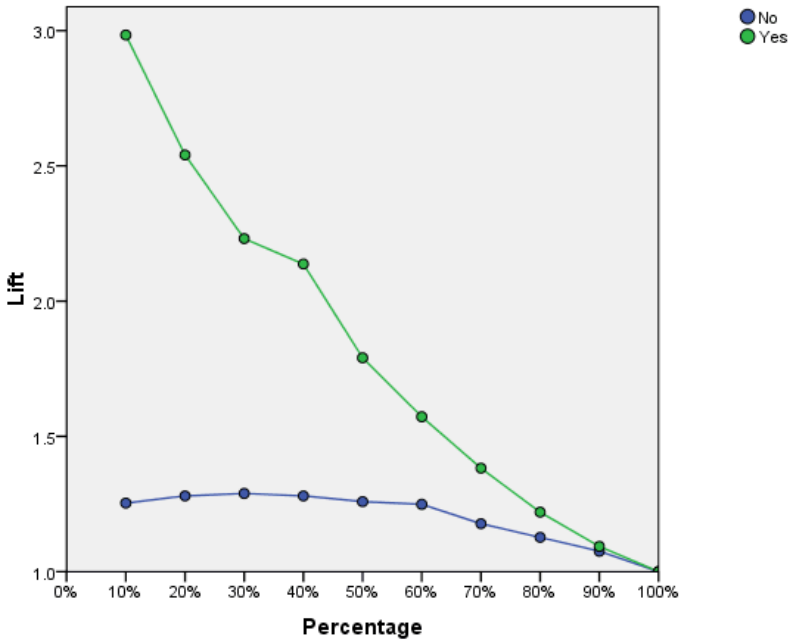
که ۱۰٪ اولیه آنها، حدوداً شامل ۳۰٪ از تمام مواردی باشد که در واقع در دسته YES (بدحسابان) قرار گرفته‌اند. به همین ترتیب ۲۰٪ اولیه در این ترتیب چینش می‌تواند شامل حدوداً ۵۰٪ از اعضای دسته بدحسابان باشد و همچنین ۳۰٪ اولیه شامل ۷۰٪ از بدحسابان خواهد بود و به همین ترتیب ادامه پیدا می‌کند.

چنانچه ۱۰۰٪ داده‌های ثبت شده را انتخاب نمایید، تمامی بدحسابان حاضر در مجموعه داده خود را، به‌دست خواهید آورد.

خط قطری را که مشاهده می‌نمایید یک منحنی مبنا است. به این معنا که، چنانچه شما ۱۰٪ از موارد موجود در مجموعه داده‌ای خود را به‌صورت تصادفی انتخاب نمایید، می‌بایست انتظار داشته باشید که در حدود ۱۰٪ از تمام مواردی که در واقع در دسته YES قرار می‌گیرند شامل شود، بالاتر از این منحنی مبنا نمودارها به‌صورتی هستند که هر چه قدر شما درصد بیشتری را انتخاب نمایید، درصد بزرگتری از موارد موجود در دسته موردنظرتان را خواهید داشت. می‌توانید از این نمودار برای تعیین شرط توقف در فرایند طبقه‌بندی استفاده نمایید. این کار را با انتخاب درصد متناظر با میزانی که می‌خواهیم در دست داشته باشیم انجام می‌دهیم و سپس این درصد را با مقدار شرط توقف موردنیاز تأمین می‌نماییم.

آن چیزی که میزان موردنظری را که می‌خواهیم به دست آوریم مشخص می‌نماید، هزینه خطاهای نوع اول و نوع دوم است. به این معنا که هزینه قرار دادن یک مشتری بدحساب در گروه مشتریان خوش‌حساب چه میزان می‌باشد؟! (خطای نوع اول) و یا این که هزینه قرار دادن یک مشتری خوش‌حساب در گروه مشتریان بدحساب چه میزان است؟! (خطای نوع دوم). چنانچه بدحسابی یک مسئله و دغدغه اساسی برای شما باشد، می‌بایست میزان خطای نوع اول را کاهش دهید. برروی نمودار Cumulative gain chart می‌توان با ردکردن تقاضای وام ۴۰٪ افراد ابتدایی که در گروه YES قرار گرفته‌اند و در حدود ۹۰٪ مشتریان بدحساب احتمالی را که دربرمی‌گیرند، میزان خطای نوع اول را به میزان زیادی کاهش داد. چنانچه افزایش میزان مشتریان بانک برای شما دارای اهمیت بیشتری باشد، می‌بایست میزان خطای نوع دوم خود را کاهش دهید، بدین‌ترتیب، برروی نمودار می‌توان با رد کردن ۱۰٪ اولیه که ۳۰٪ از بدحسابان را در خود جای می‌دهند، مقدار بیشتری از مشتریان خود را پاسخگو بود. معمولاً هر دوی این شرایط خطرناک و نامناسب می‌باشند، بنابراین شما، می‌بایست سیاست تصمیم‌سازی مناسبی

را برای دسته‌بندی مشتریان اتخاذ نمایید که مناسب‌ترین ترکیب از میان دو عنصر عنوان شده را به دست دهد.




شکل ۳-۲۴

نمودار lift chart از نمودار Cumulative gains chart حاصل می‌شود. برای هر یک از منحنی‌ها مقادیر موجود برروی نمودار Y متناظر است با نسبت مقادیر موجود برروی نمودار Cumulative gain chart، بنابراین میزان Lift برای ۱۰٪ از دسته YES برابر است با

$$\frac{\٪۳۰}{\٪۱۰} = ۳$$

این نمودار شیوه جدیدی را برای نگاه به اطلاعات موجود در نمودار cumulative gains chart فراهم می‌سازد.

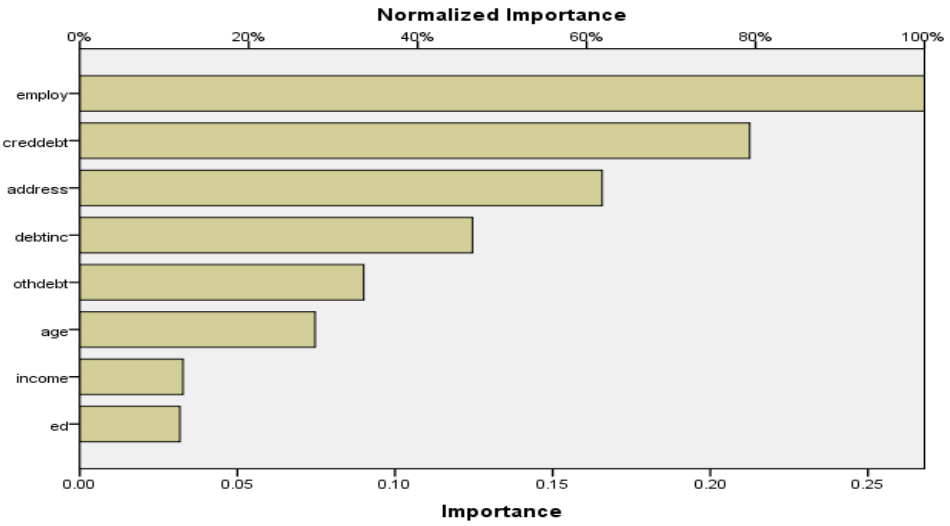
**نکته:** نمودارهای Lift chart و Cumulative gains chart براساس نمونه‌هایی مرکب از نمونه‌های آموزشی و آزمایشی ترسیم می‌شوند.

## اهمیت متغیرهای مستقل

| | Importance | Normalized Importance |
|---|---|---|
| Level of education | .032 | 11.9% |
| Age in years | .075 | 27.9% |
| Years with current employer | .268 | 100.0% |
| Years at current address | .166 | 61.8% |
| Household income in thousands | .033 | 12.2% |
| Debt to income ratio (x100) | .125 | 46.5% |
| Credit card debt in thousands | .213 | 79.3% |
| Other debt in thousands | .090 | 33.6% |

شکل ۳-۲۵

اهمیت متغیرهای مستقل در تشخیص این نکته است که به چه میزان مقادیر پیش‌بینی شده توسط شبکه، با تغییر مقادیر متغیر مستقل، تغییر می‌نماید. نرمال‌سازی این اهمیت بسیار ساده است و با تقسیم مقادیر اهمیت بر بزرگترین مقدار آن حاصل می‌شود و به‌صورت درصد بیان می‌گردد.




شکل ۳-۲۶

نمودار Importance chart یک نمودار میله‌ای ساده حاصل از مقادیر موجود در جدول importance می‌باشد که این مقادیر به‌صورت نزولی مرتب شده‌اند. به نظر می‌رسد که متغیرهای مرتبط با ویژگی‌هایی که نشان‌دهنده ثبات مشتریان می‌باشند، از قبیل وضعیت اشتغال، آدرس و همچنین میزان بدهی آنان، بیشترین تأثیر را برروی این که شبکه چگونه آنان را طبقه‌بندی نماید، دارند. آنچه را که نمی‌توان صراحتاً عنوان داشت، جهت "direction" روابط موجود میان این متغیرها و احتمال پیش‌بینی شده در موارد بدحسابی است. ممکن است این‌گونه به نظر برسد که مقدار زیاد بدهی، نشان‌دهنده احتمال بدحسابی بیشتر است، اما مطمئن باشید که در صورت استفاده از شبکه، با مدل‌هایی روبه‌رو خواهید بود که دارای پارامترهای تفسیرپذیر راحت‌تری باشند.

## خلاصه

برای استفاده از فرایند چند لایه پرسپترون می‌بایست شبکه‌ای را برای پیش‌بینی احتمال این که مشتری دریافت وام، بدحسابی خواهد کرد یا خیر تشکیل دهید. نتایج حاصل از این مدل با نتایجی که از روش رگرسیونی و یا آنالیزهای دیگر به‌دست می‌آید، قابل مقایسه می‌باشند، بنابراین می‌توانید مطمئن باشید که روابط به وجود آمده میان داده‌ها در این مدل، با روابط حاصل در مدل‌های دیگر تفاوتی نداشته و قابل دسترسی می‌باشند. بدین‌ترتیب می‌توانید از آن برای بررسی بیشتر طبیعت روابط موجود میان متغیرهای مستقل و وابسته استفاده نمایید.

# استفاده از پرسپترون چند لایه به‌منظور محاسبه هزینه‌های درمانی و مدت زمان بستری بیماران

یک سیستم بیمارستانی علاقه‌مند به ایجاد ساختاری از هزینه‌ها و مدت زمان بستری بیمارانی است که جهت انجام درمان پذیرش می‌شوند، این بیماران مبتلا به نارسایی‌های قلبی نظیر حمله قلبی و.... می‌باشند. دسترسی به نتایج این محاسبات به‌گونه‌ای قابل اطمینان و مناسب به بیمارستان این اجازه را می‌دهد که به مدیریت مناسبی از فضاها و تخت‌های آماده برای بیماران دست یابند.

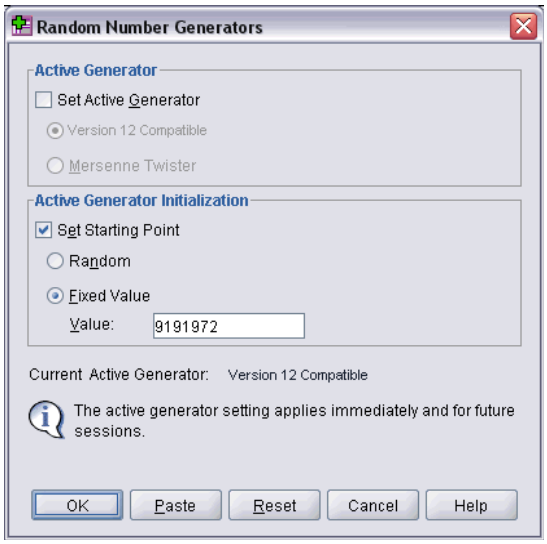
داده‌های موجود در فایل Patient-los.sav شامل اطلاعات پیرامون بیمارانی است که جهت درمان نارسایی قلبی نیاز به مراقبت‌های درمانی دارند. برای اطلاعات بیشتر، می‌توانید به فایل نمونه‌های در پیوست A مراجعه نمایید. از فرایند پرسپترون چند لایه جهت ساخت شبکه‌ای برای پیش‌بینی هزینه‌ها و طول مدت بستری بیماران استفاده نمایید.

## آماده‌سازی داده‌ها جهت انجام تحلیل‌ها

ایجاد دسته‌بندی‌های تصادفی این اجازه را به شما می‌دهد که شرایط انجام تحلیل‌ها را به دقت و درستی فراهم آورید.

- برای ایجاد دسته‌بندی‌های تصادفی، از منو، انتخاب نمایید:

Transform→Random number generators…



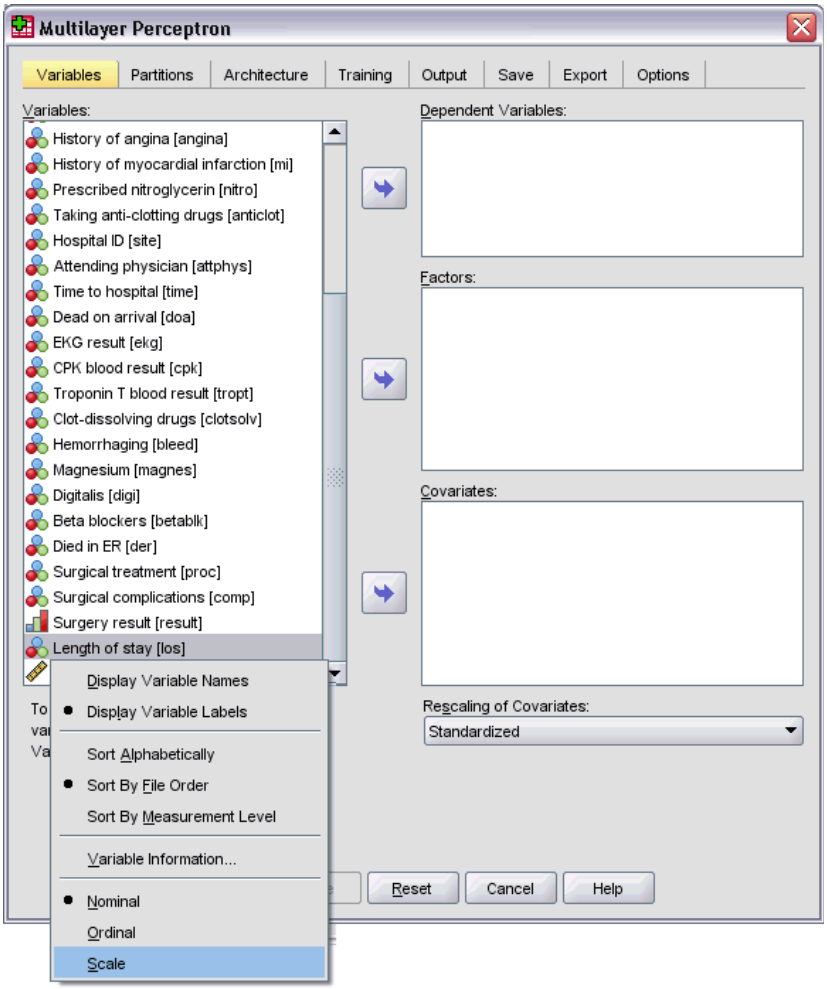

شکل ۳-۲۷

- گزینه set starting point را انتخاب نمایید.
- گزینه Fixed value را انتخاب نموده و عدد ۹۱۹۱۹۷۲ را برای مقدار خواسته شده تایپ نمایید.
- گزینه ok را کلیک کنید.

## آغاز آنالیزها

برای راه‌اندازی آنالیزهای پرسپترون چندلایه، می‌بایست از منو انتخاب نمایید:

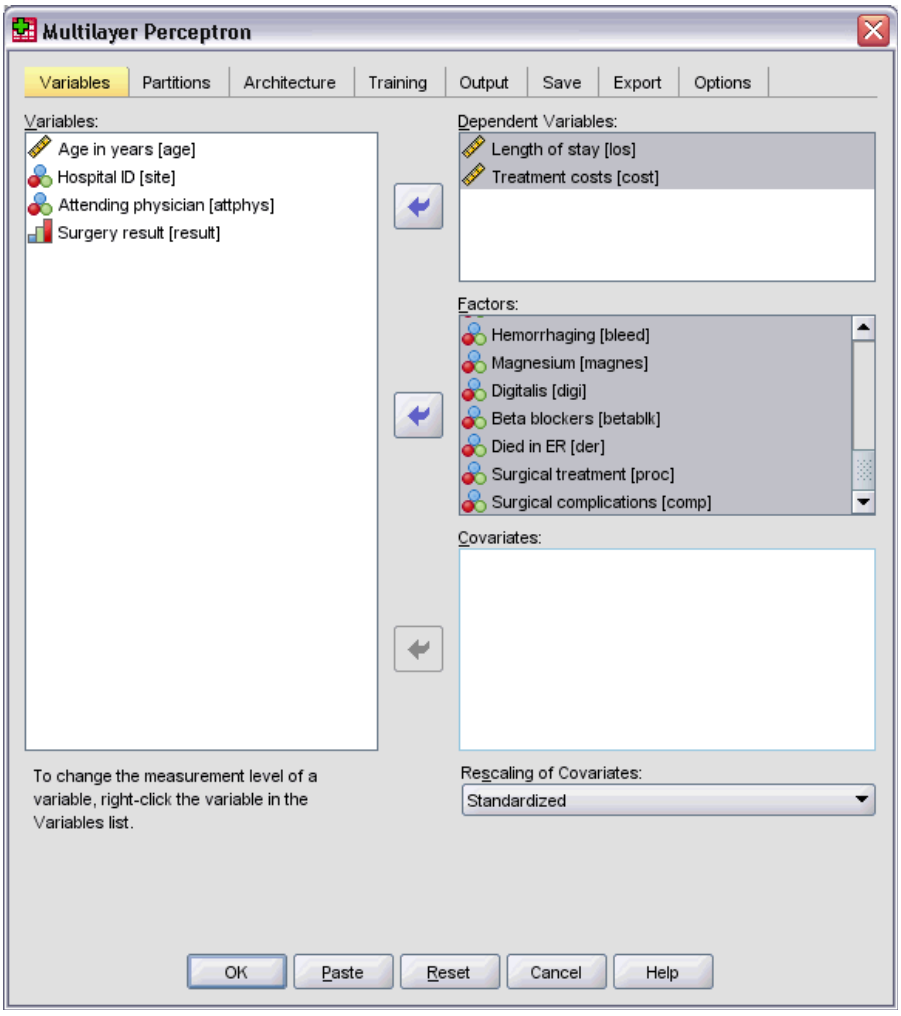
Analyze→Neural Network→Multilayer Perceptron




شکل ۳-۲۸

طول مدت بستری بیماران [los] دارای سطح اندازه‌گیری ترتیبی است، اما در این موردنیاز است که شبکه با آن به‌عنوان یک مقیاس رفتار نماید.

- برروی عبارت Length of stay راست کلیک کرده و از روی صفحه باز شده عبارت scale را انتخاب نمایید.

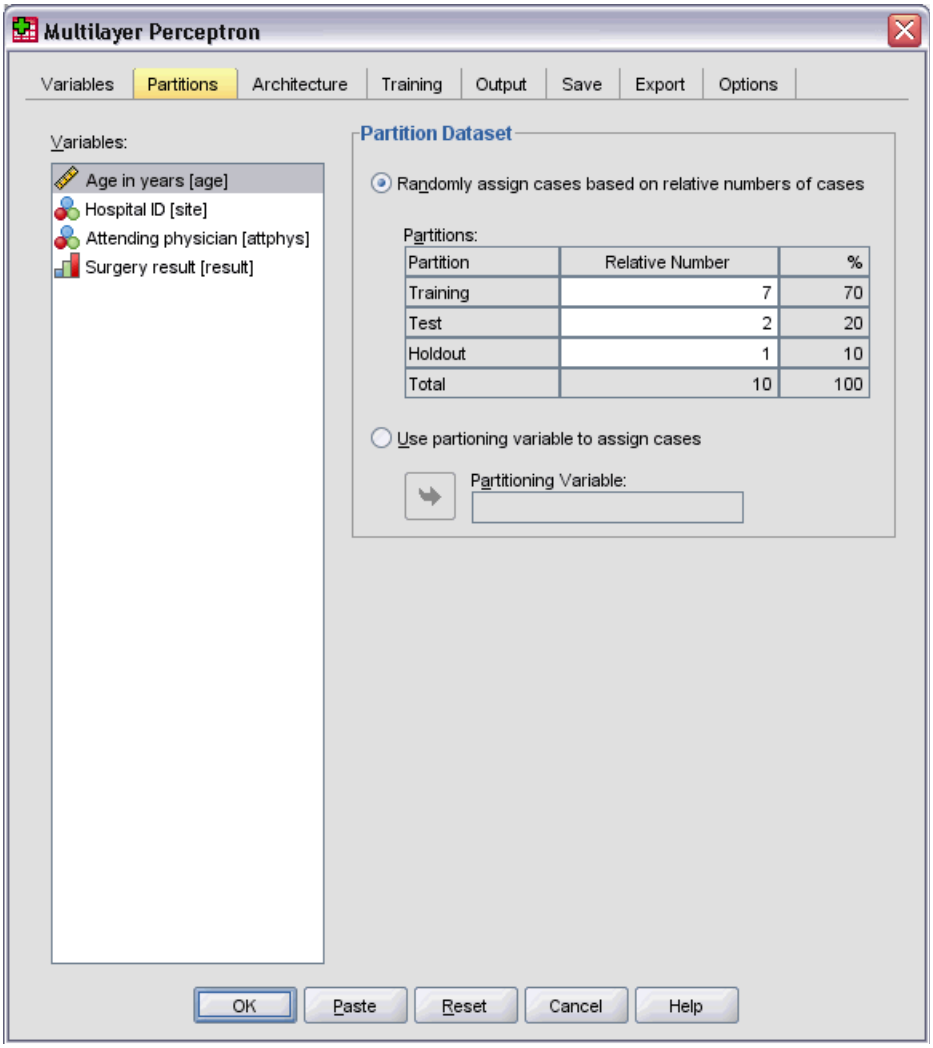



شکل ۳-۲۹

- عبارات length of stay و treatment costs را به‌عنوان متغیرهای وابسته انتخاب کنید.
- عبارات age category تا talking anti-clotting drugs و time to hospital تا surgical complication را به‌عنوان فاکتورها انتخاب نمایید. برای حصول اطمینان از قابلیت تکرارپذیری نتایج مدل،

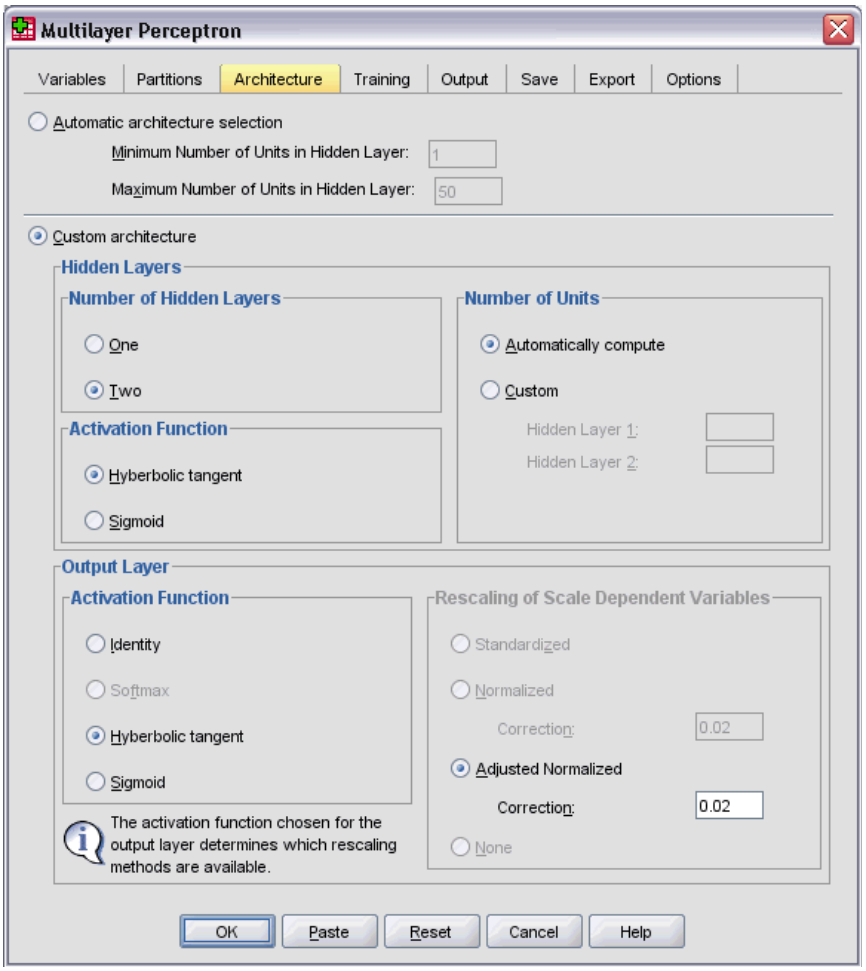
مطمئن شوید که ترتیب متغیرهای موجود در لیست factor درست قرار گرفته است. برای انجام این کار و مرتب نمودن ترتیب این عوامل بهتر است با کلیک برروی آنها و سپس استفاده از دکمه‌های روی ماوس جهت جابه‌جایی و کشیدن گزینه‌ها استفاده نمایید. جابه‌جایی ترتیب قرار گرفتن متغیرها به‌صورت متناوب به شما در ارزیابی میزان پایداری راه‌حل ارائه شده کمک می‌نماید.

- برروی گزینه partition کلیک کنید.




شکل ۳-۳۰

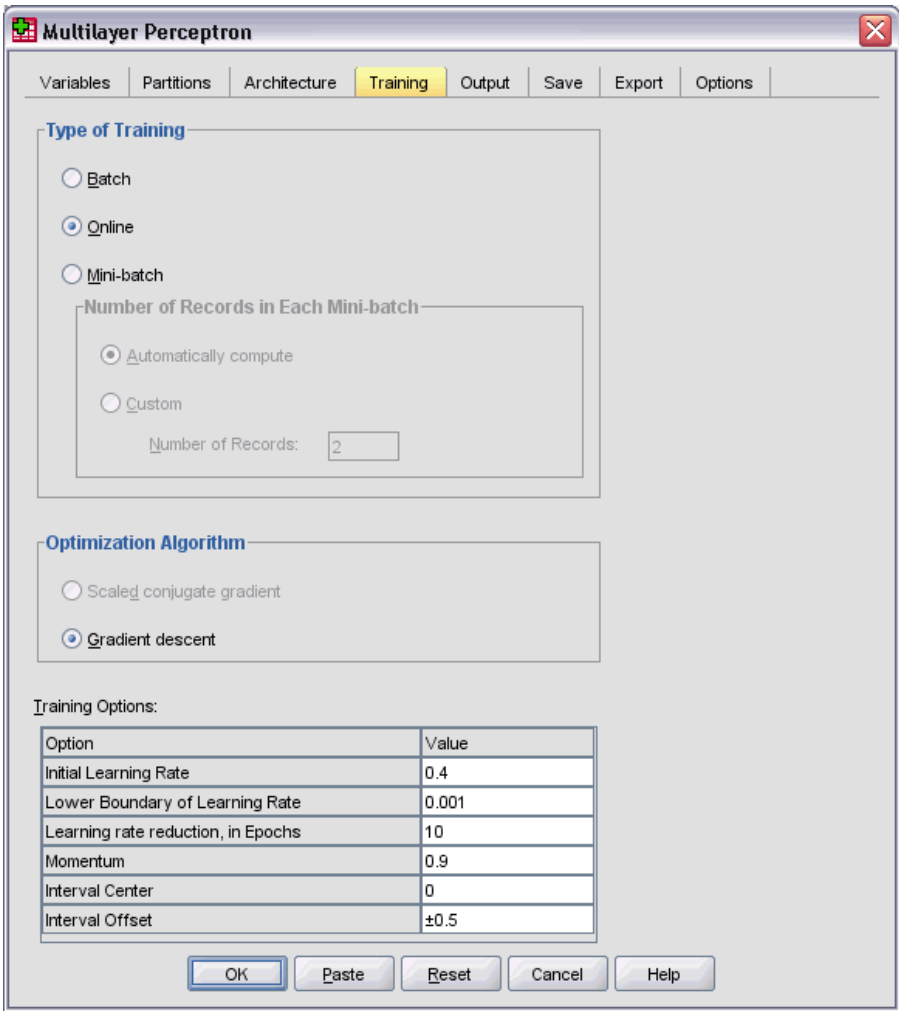
- در قسمت relative number و در مقابل عنوان نمونه آزمایشی عدد ۲ را تایپ کنید.
- در قسمت relative number و در مقابل عنوان نمونه جدا از هم نگهداری شده (holdout) عدد ۱ را تایپ کنید.
- برروی گزینه architecture کلیک نمایید.




شکل ۳-۳۱

- گزینه custom architecture را انتخاب کنید.
- گزینه two را به‌عنوان تعداد لایه پنهان انتخاب نمایید.

- گزینه hyperbolic tangent را به‌عنوان تابع فعال‌ساز لایه خروجی انتخاب نمایید. توجه‌داشته باشید که این تابع فعال‌ساز، به‌صورت خودکار از روش مقیاس‌بندی مجدد بر متغیرهای وابسته استفاده می‌نماید تا به داده‌های نرمال شده دست یابیم.
- برروی گزینه training کلیک کنید.

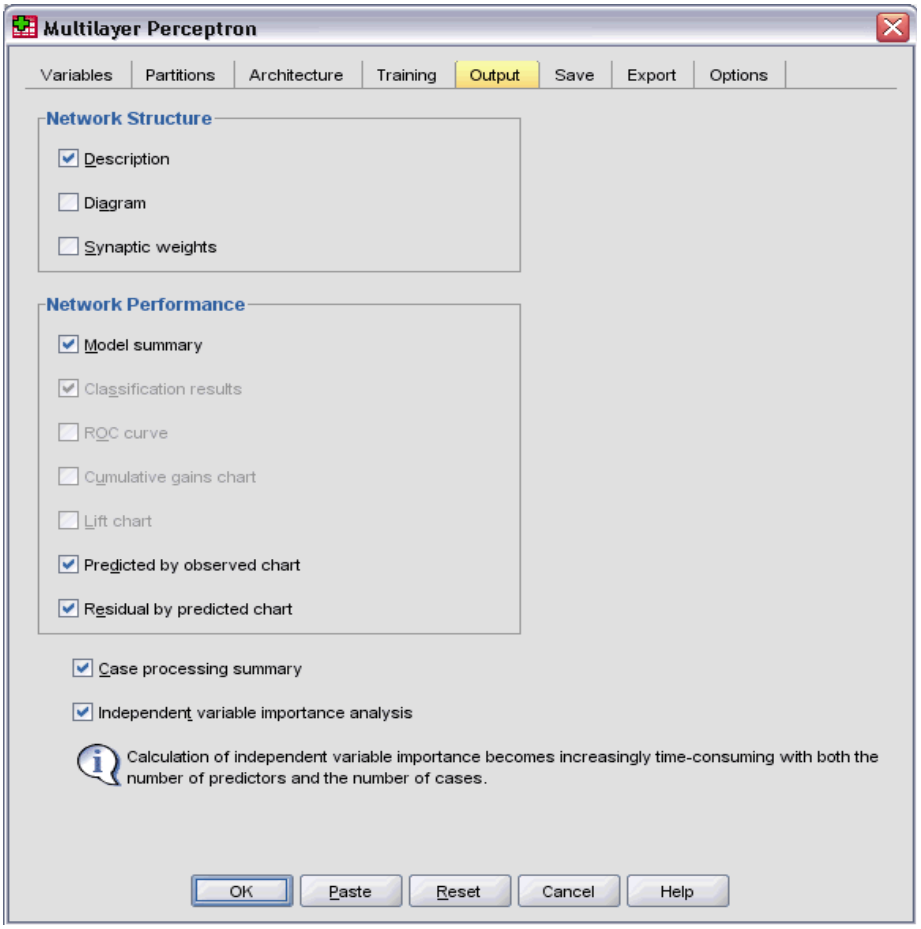



*شکل ۳-۳۲*

- در میان گزینه‌های موجود در بخش type of training گزینه Online را انتخاب نمایید. این نوع آموزش در مورد مجموعه داده‌های بزرگتر با پیش‌بینی‌کننده‌های همبسته عملکرد

مناسبتری دارد. توجه داشته باشید که از طریق الگوریتم بهینه‌سازی به‌صورت خودکار شیب به‌صورت نزولی خواهد بود.

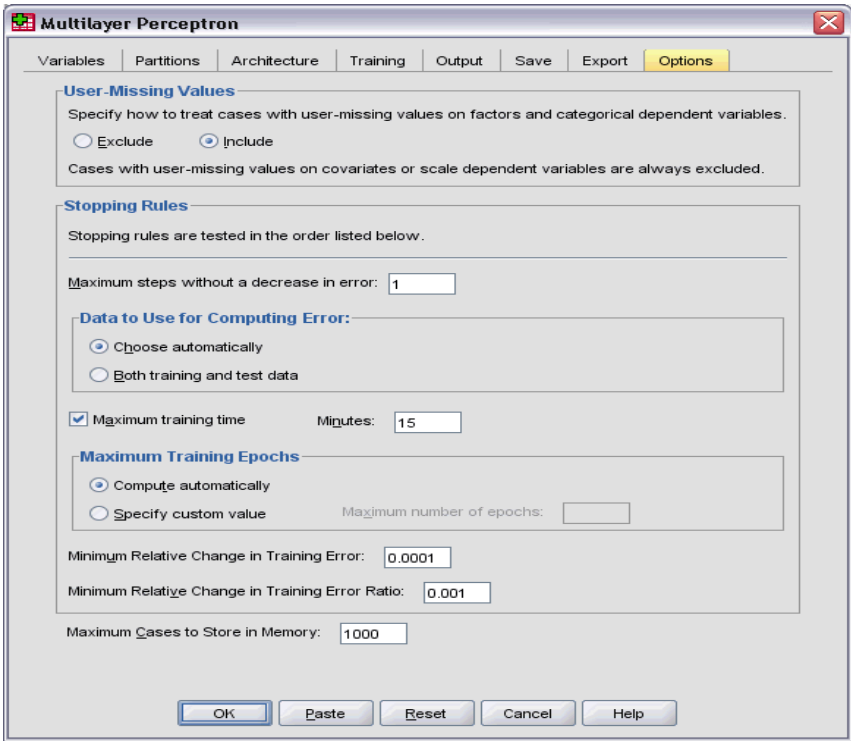
- برروی گزینه Output کلیک کنید.




*شکل ۳-۳۳*

- گزینه diagram را انتخاب ننمایید. میزان ورودی‌ها بسیار زیاد بوده و در نتیجه دیاگرام سنگین و بزرگ خواهد بود.
- گزینه‌های Predicted by observed chart و Residual by predicted را در قسمت network performance انتخاب کنید. گزینه‌های دیگر از جمله classification، ROC curve، cumulative gain chart و lift chart از آنجایی که هیچ متغیر وابسته‌ای به‌صورت مطلق (اسمی یا وصفی)، شناخته نشده‌اند، در دسترس نمی‌باشند.

- گزینه Independent variable analysis را انتخاب نمایید.
- برروی گزینه option کلیک کنید.




*شکل ۳-۳۴*

- در قسمت user-missing values ، گزینه Include را انتخاب نمایید، با این کار متغیرهای ناپیدا نیز در نظر گرفته می‌شوند. بیمارانی که تحت عمل جراحی قرار نمی‌گیرند دارای این مقادیر گم شده در متغیر surgical complications هستند، این عمل برای اطمینان در این مورد است که این بیماران نیز در آنالیزها در نظر گرفته می‌شوند.
- برروی گزینه ok کلیک کنید.

## اعلام خطرها


*شکل ۳-۳۵*

جدول Warnings، نشان می‌دهد که متغیرهای doa و der در نمونه آموزشی ثابت هستند. بیمارانی که در بدو ورود به بیمارستان مرده‌اند و یا در اتاق‌های اورژانس می‌میرند در length of stay دارای مقادیر گم شده (user-missing values) هستند.

از آنجایی که در این آنالیزها، طول مدت بستری بیماران (length of stay) به‌عنوان یک متغیر مقیاسی در نظر گرفته‌ایم و مواردی که دارای مقادیر گم شده می‌باشند در این مقیاس جای نمی‌گیرند، تنها بیمارانی که پس از گذر از اتاق اورژانس زنده هستند را در داخل این متغیر قرار می‌دهیم.

## خلاصه فرایند

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 5647 | 70.6% |
| | Testing | 1570 | 19.6% |
| | Holdout | 781 | 9.8% |
| Valid | | 7998 | 100.0% |
| Excluded | | 2002 | |
| Total | | 10000 | |

*شکل ۳-۳۶*

این جدول نشان می‌دهد که، ۵۶۴۷ مورد به‌عنوان نمونه آموزشی، ۱۵۷۰ مورد نمونه آزمایشی و ۷۸۱ مورد نمونه جدانگه داشته شده در نظر گرفته شده‌اند. همان‌طورکه مشاهده می‌کنید، ۲۰۰۲ مورد از کل موارد در دسترس از آنالیزها خارج گردیده‌اند، که در واقع بیمارانی هستند که در طی انتقال به بیمارستان و یا در اتاق اورژانس فوت کرده‌اند.

## اطلاعات شبکه

| | | | |
|---|---|---|---|
| Input Layer | Factors | 1 | Age category |
| | | 2 | Gender |
| | | 3 | History of diabetes |
| | | 4 | Blood pressure |
| | | 5 | Smoker |
| | | 6 | Cholesterol |
| | | 7 | Physically active |
| | | 8 | Obesity |
| | | 9 | History of angina |
| | | 10 | History of myocardial infarction |
| | | 11 | Prescribed nitroglycerin |
| | | 12 | Taking anti-clotting drugs |
| | | 13 | Time to hospital |
| | | 14 | EKG result |
| | | 15 | CPK blood result |
| | | 16 | Troponin T blood result |
| | | 17 | Clot-dissolving drugs |
| | | 18 | Hemorrhaging |
| | | 19 | Magnesium |
| | | 20 | Digitalis |
| | | 21 | Beta blockers |
| | | 22 | Surgical treatment |
| | | 23 | Surgical complications |
| | Number of Units[a] | | 63 |
| Hidden Layer(s) | Number of Hidden Layers | | 2 |
| | Number of Units in Hidden Layer 1[a] | | 12 |
| | Number of Units in Hidden Layer 2[a] | | 9 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | Length of stay |
| | | 2 | Treatment costs |
| | Number of Units | | 2 |
| | Rescaling Method for Scale Dependents | | Adjusted Normalized |
| | Activation Function | | Hyperbolic tangent |
| | Error Function | | Sum of Squares |

a.Excluding the bias unit

*شکل ۳-۳۷*

جدول اطلاعات شبکه، اطلاعات مربوط به شبکه عصبی را نمایش می‌دهد و برای حصول اطمینان از این که شبکه عمل طبقه‌بندی را به درستی انجام داده است، مفید می‌باشد. در اینجا، مهم است که به موارد زیر به‌صورت خاص توجه نمایید:

- تعداد واحدهای موجود در لایه ورودی برابر با مجموع تعداد سطوح ضرایب می‌باشد (در اینجا کوواریانس وجود ندارد).
- دو لایه پنهان درخواست شده و فرایند ۱۲ واحد را در لایه اول و ۹ واحد را در لایه دوم پنهان، قرار داده است.
- برای هر یک از متغیرهای وابسته به مقیاس یک واحد خروجی جدا و مستقل ایجاد گردیده است. این واحدها توسط روش نرمال شده از نو مقیاس شده‌اند، که در آن برای لایه خروجی نیاز به تابع فعال‌ساز تانژانت هیبربولیک داریم.

- در جدول، میزان مجموع مربعات خطاها نیز آورده شده است، زیرا که متغیرهای وابسته، مقیاس هستند.

## خلاصه مدل

| | | | |
|---|---|---|---|
| Training | Sum of Squares Error | | 91.812 |
| | Average Overall Relative Error | | .083 |
| | Relative Error for Scale Dependents | Length of stay | .131 |
| | | Treatment costs | .033 |
| | Stopping Rule Used | | 1 consecutive step (s) with no decrease in error[a] |
| | Training Time | | 00:00:18.055 |
| Testing | Sum of Squares Error | | 26.798 |
| | Average Overall Relative Error | | .088 |
| | Relative Error for Scale Dependents | Length of stay | .141 |
| | | Treatment costs | .033 |
| Holdout | Average Overall Relative Error | | .099 |
| | Relative Error for Scale Dependents | Length of stay | .154 |
| | | Treatment costs | .041 |

a.Error computations are based on the testing sample.

*شکل ۳-۳۹*

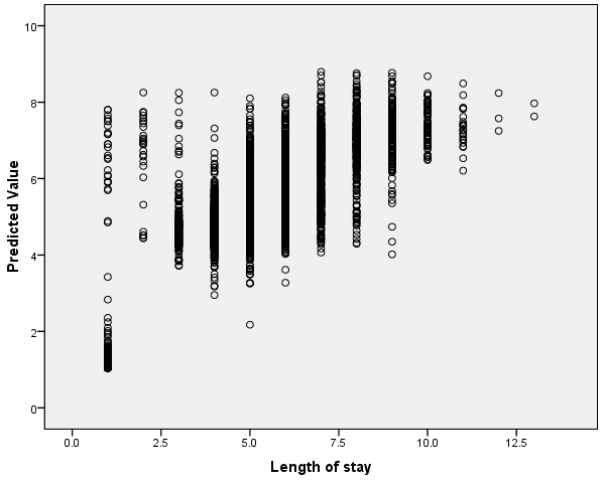
- در این جدول مجموعه مربعات خطاها از آنجایی که لایه خروجی دارای متغیرهای وابسته به مقیاس می‌باشند، نشان داده شده است. این همان تابع خطایی است که شبکه تلاش دارد در طی عملیات آموزش آن را به کمترین میزان خود برساند. توجه داشته باشید که مقادیر مربوط به مجموع مربعات و مابقی خطاها که در جدول آورده شده است، برای مقادیر از نو مقیاس شده متغیرهای وابسته، محاسبه گردیده‌اند.
- خطای متناظر با هر یک از متغیرهای وابسته به مقیاس، حاصل نسبتی است از مجموع مربعات خطاهای متغیر وابسته، به مجموع مربعات خطاهای مدل تهی (null). که در آن از مقدار میانگین متغیر وابسته به‌عنوان مقدار پیش‌بینی برای هر یک از موارد استفاده می‌شود. مشاهده می‌شود که میزان خطای انجام شده در پیش‌بینی مدت زمان بستری بیماران (length of stay) نسبت به پیش‌بینی هزینه‌های درمان (treatment costs) آنان، بیشتر می‌باشد.
- میانگین کل خطا نسبتی است از مجموع مربعات خطای تمام متغیرهای وابسته، به مجموع مربعات خطای موجود درمدل تهی، که در آن از مقادیر میانگین متغیرهای وابسته به‌عنوان

مقادیر پیش‌بینی شده برای هر یک از موارد استفاده می‌گردد. در این مثال، تصادفاً، میزان میانگین کل خطاها نزدیک به مقدار میانگین خطاهای نسبی است، اما باید این نکته را در نظر داشت که در تمامی موارد این قاعده وجود نخواهد داشت.

میزان میانگین کل خطای نسبی و خطاهای نسبی تاحدودی در نمونه‌های آزمایشی، آموزشی و نمونه‌های جدا نگه داشته شده ، ثابت می‌باشد. این مسئله این اطمینان را به همراه دارد که در مدل آموزش اضافی رخ نمی‌دهد و میزان خطاهایی که در آینده توسط شبکه ثبت می‌گردد، به میزان خطاهای گزارش شده در این جدول نزدیک خواهد بود.

- از آنجایی که میزان خطا پس از گذشت یک مرحله از الگوریتم تخمین کاهش نیافته، این الگوریتم متوقف گردیده است.

## جدول پیش‌بینی براساس مشاهده (Predicted-by-Observed Charts)



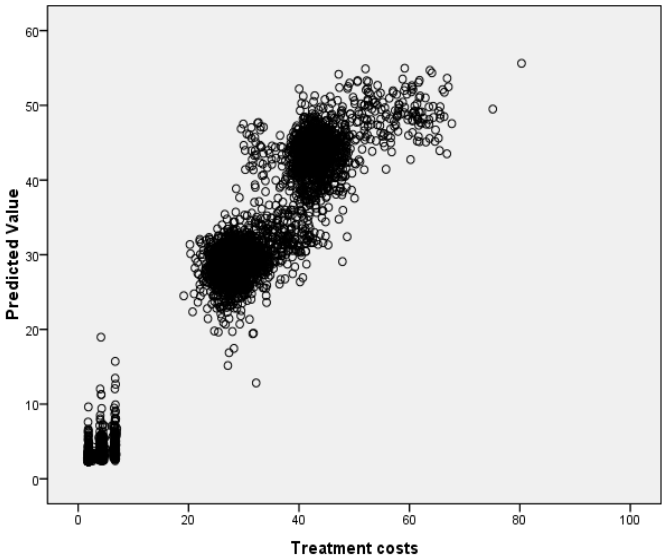

*شکل ۳-۴۰*

جدول پیش‌بینی براساس مشاهده، در موارد مربوط به متغیرهای وابسته به مقیاس، نمودار پراکندگی است متشکل از مقادیر پیش‌بینی شده و مقادیر مشاهده شده برای نمونه‌های آزمایشی و آموزشی به‌طوری‌که مقادیر پیش‌بینی شده برروی محور Y و مقادیر مشاهده شده برروی محور x قرار می‌گیرند.

در حالت ایده‌آل، این مقادیر می‌بایست به‌طور تقریبی در طول خط ۴۵ درجه که از مبدأ آغاز می‌گردد، قرار گیرند. در این جدول، خطوط عمودی حاصل از نقاط مربوط به تعداد روزهای بستری بیماران می‌باشد.

به جدول نگاه کنید، همان‌طورکه مشاهده می‌شود، شبکه عملکرد مناسبی را در پیش‌بینی مدت زمان بستری بیماران داشته است. شکل و روند کلی نمودار نشان می‌دهد که مقادیر دور از خط ۴۵ درجه ایده‌آل قرار گرفته‌اند و این نکته بدان معناست که پیش‌بینی‌ها در مواردی که طول مدت بستری کمتر از ۵ روز ثبت گردیده، بیش از حد واقعی و نامناسب تخمین زده شده است و در مواردی که مقدار ثبت شده بیش از ۶ روز بوده، میزان پیش‌بینی کمتر از حد واقعی مدت زمان بستری صورت پذیرفته است.

گروه بیمارانی که در قسمت پایین و سمت چپ نمودار قرار گرفته‌اند، بیمارانی هستند که به احتمال زیاد نیازی به عمل جراحی ندارند، همچنین دسته دیگری از بیماران نیز در قسمت بالایی و سمت چپ نمودار واقع شده‌اند که در مورد آنان تعداد روزهای بستری یک تا سه روز مشاهده شده درصورتی‌که مقدار پیش‌بینی شده این مؤلفه در مورد آنان بسیار بیشتر از این مقادیر بوده است. به نظر می‌رسد این موارد، بیمارانی هستند که پس از انجام عمل جراحی در بیمارستان فوت کرده‌اند.



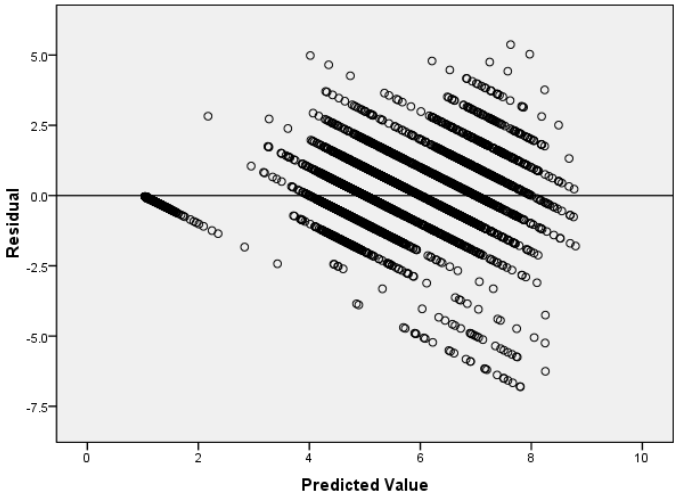

*شکل ۳-۴۱*

آن‌گونه که مشاهده می‌شود، شبکه در پیش‌بینی هزینه‌های درمان نیز عملکرد مناسبی داشته است و سه دسته مقدماتی برای بیماران در نظر گرفته شده است.

- در قسمت پایینی سمت چپ بیمارانی هستند که تحت عمل جراحی قرار نگرفته اند ، هزینه درمان آنان نسبتاً کم بوده و با توجه به نوع دارویی که در اتاق اورژانس برای آنان تجویز شده است متفاوت می‌باشند.
- دسته دیگری از بیماران وجود دارند که هزینه درمان آنها در حدود ۳۰۰۰۰ دلار است.این بیماران تحت درمان PTCA (بازکردن عروق کرونری به وسیله بالون) قرار گرفته‌اند.
- دسته سوم بیمارانی هستند که هزینه درمان آنان در حدود ۴۰۰۰۰ دلار بوده است. این بیماران تحت عمل جراحی باز برای ترمیم عروق کرونری قلب (CABG)قرار گرفته‌اند. این عمل گرانتر از درمان PTCA بوده و بیماران دوران نقاهت بیشتری را پس از انجام عمل جراحی ، پشت سر می گذارند،که این مسئله باعث افزایش هزینه ها می گردد.

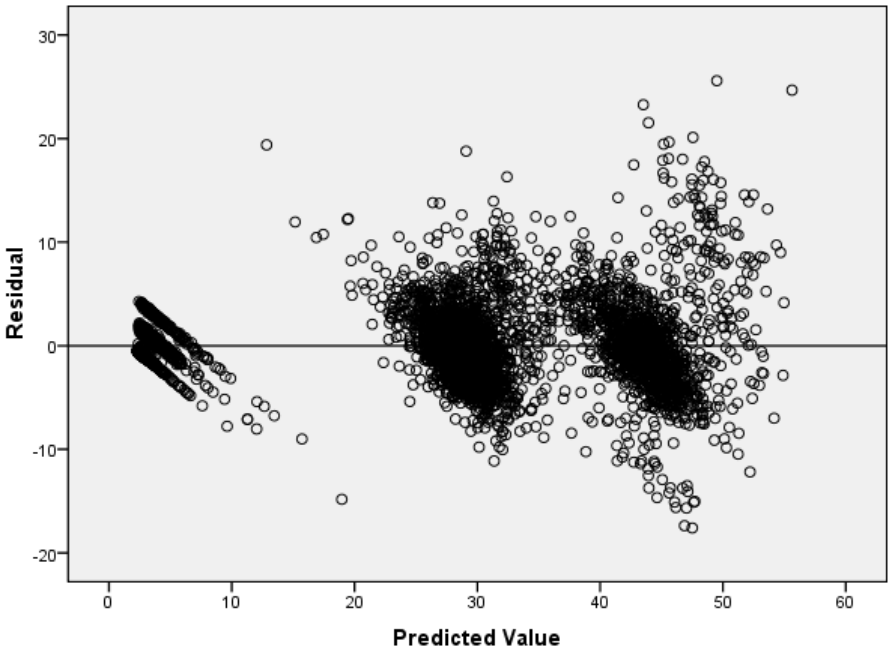
در ضمن مواردی نیز وجود دارند که هزینه درمان آنها متجاوز از ۵۰۰۰۰ دلار بوده که شبکه عملکرد مناسبی در پیش بینی آنها ندارد.این بیماران افرادی هستند که در طی عمل جراحی با مشکلات و مسائل پیچیده ای روبه رو می شوند که باعث افزایش هزینه عمل جراحی و مدت زمان بستری آنان در بیمارستان می گردد.

## Residual by predicted chart




شکل ۳-۴۱

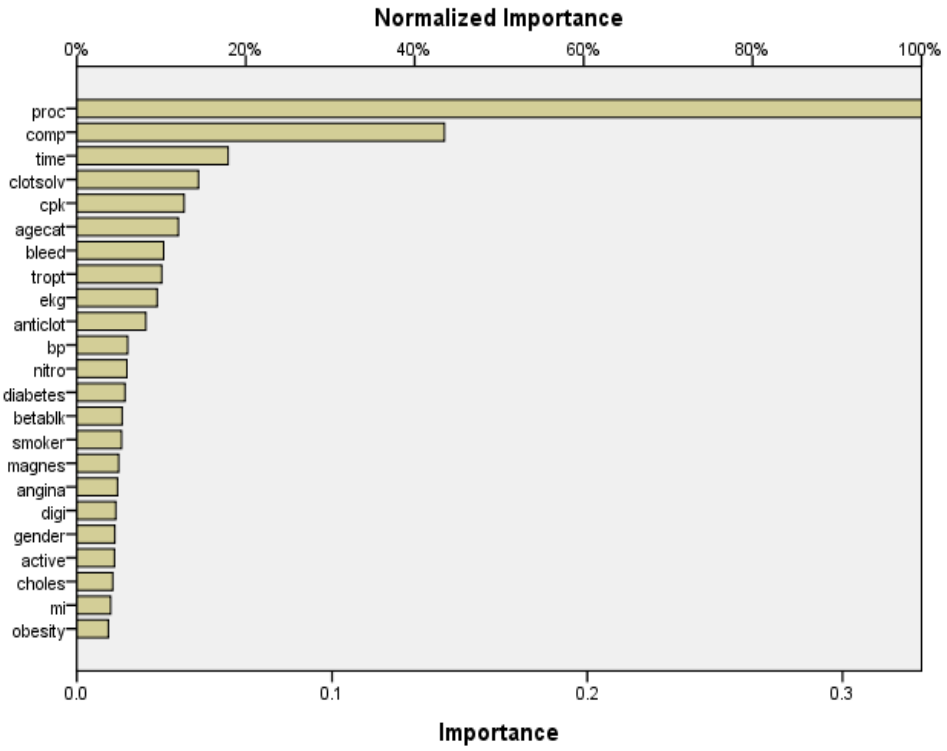
این نمودار، یک نمودار پراکندگی است که در آن مقادیر باقی مانده (برابر با مقدار مشاهده شده منهای مقدار پیش‌بینی شده) برروی محور Y و مقادیر پیش‌بینی شده برروی محور x آن قرار می‌گیرند. هریک از خطوط اریب موجود در این نمودار متناظر با یکی از خطوط عمودی موجود در نمودار پیش‌بینی به‌وسیله مشاهده است و همچنین به‌صورت واضح می‌توان انتقال و توالی از پیش‌بینی‌های بیش از اندازه به پیش‌بینی‌های کمتر از میزان واقعی را تحت‌تأثیر افزایش مدت مشاهده شده زمان بستری بیماران نشان داد.




*شکل ۳-۴۲*

در مورد مؤلفه هزینه‌های درمان (Treatment costs)، همچون مؤلفه طول مدت بستری بیماران (length of stay)، در این نمودار با افزایش میزان هزینه‌های مشاهده شده درمان بیماران، انتقالی از پیش‌بینی‌های مناسب انجام شده به پیش‌بینی‌های کمتر از میزان واقعی، مشاهده می‌گردد. بیمارانی که در طی درمان CABG دچار مشکل شده و عمل آنها با پیچیدگی همراه شده است به وضوح قابل تشخیص می‌باشند، با این وجود تشخیص بیمارانی که در طی درمان PTCA با مشکلاتی از این قبیل روبه‌رو گردیده‌اند از این نیز ساده‌تر است. بدین شکل که این بیماران به‌صورت زیرگروهی کوچک در سمت راست و بالای گروه بیمارانی که تحت

درمان PTCA قرار گرفته‌اند دیده می‌شوند (بیمارانی که هزینه درمانشان بیش از ۳۰۰۰۰ دلار بوده و برروی محور x می‌توان آنان را تشخیص داد.)




*شکل ۳-۴۳*

این نمودار نشان می‌دهد که نتایج تحت‌تأثیر چگونگی و کیفیت انجام عمل جراحی قرار دارند. تمام موارد از قبیل این که آیا در طی عمل جراحی با مشکل فنی روبه‌رو شده‌ایم و یا این که عوامل دیگر در نتیجه و کیفیت انجام عمل تأثیر گذاشته‌اند یا خیر، در این نمودار قابل تشخیص است.

اهمیت فرایند عمل جراحی و تأثیر آن بر هزینه‌های درمان در این نمودار به روشنی قابل مشاهده است. اما در مورد مؤلفه مدت زمان بستری بیماران این تأثیر به دلایلی کم‌رنگ‌تر می‌باشد. با این وجود تأثیر وقوع مشکلات فنی در طی عمل جراحی برروی مدت زمان بستری

بیماران در عالم واقع با مشاهده بیمارانی که به زمان بیشتری جهت طی دوران نقاهت نیاز دارند، مشخص می‌گردد.

## خلاصه

شبکه در مورد بیماران عادی، عملکرد مناسبی را از خود نشان می‌دهد، اما در مورد بیمارانی که پس از عمل جراحی فوت می‌شوند، ناموفق عمل می‌کند. برای رفع این مشکل یک راه‌حل می‌تواند ایجاد شبکه‌های چندگانه باشد که در آن یک شبکه نتایج بیماران را فارغ از این که زنده بمانند یا فوت کنند، پیش‌بینی می‌نماید و پس از آن شبکه‌ای دیگر به پیش‌بینی هزینه‌های حاصل از درمان و طول مدت بستری بیماران می‌پردازد.

سپس شما می‌توانید با ادغام نتایج حاصل از این دو شبکه به نتایج مناسبتری دسترسی پیدا کنید. برای حل مشکل مربوط به هزینه‌های درمان و طول مدت بستری بیمارانی که در طی انجام عمل جراحی با مشکل روبه‌رو می‌شوند نیز می‌توانید از راه‌حلی مشابه‌ای استفاده نمایید.

# بخش دوم

## تابع شعاع مدار

فرایند تابع شعاع مدار (RBF) مدل پیش‌بینی‌کننده‌ای را براساس مقادیر متغیرهای پیش‌بینی، برای یک یا چند متغیر وابسته (هدف) تولید می‌نماید.

## استفاده از RBF جهت طبقه‌بندی مشتریان خدمات ارتباط از راه‌دور

یک ارائه‌کننده خدمات ارتباط از راه‌دور، اقدام به دسته‌بندی مشتریان خود براساس الگوی مصرفی آنان نموده است و بر این اساس مشتریان خود را به ۴ دسته تقسیم کرده است. چنانچه از داده‌های حاصل از آمارگیری نفوس جهت پیش‌بینی اعضای این گروه‌ها استفاده نماییم، می‌توانیم به تشخیص مشتریان بالقوه دست یابیم.

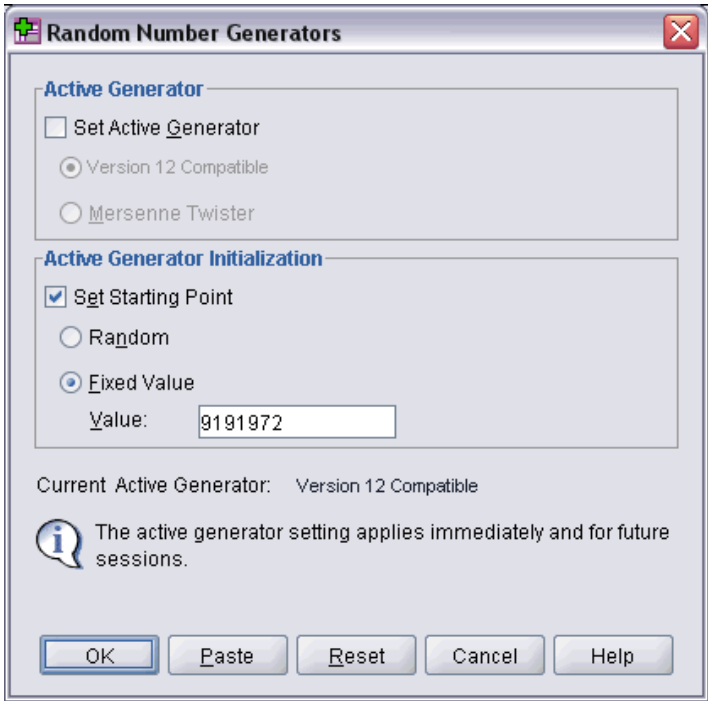
اطلاعات مربوط به مشتریانی که در حال حاضر از خدمات استفاده می‌نمایند در فایل Telco.sav موجود می‌باشد. برای اطلاعات بیشتر، می‌توانید به فایل‌های نمونه موجود در پیوست A مراجعه کنید. از فرایند تابع شعاع مدار جهت طبقه‌بندی مشتریان استفاده نمایید.

## آماده‌سازی داده‌ها جهت آغاز آنالیزها

فراهم نمودن دسته‌های تصادفی به شما این امکان را می‌دهد که آنالیزها رابه دقت تکرار نمایید.

- برای فراهم نمودن یک دسته تصادفی از بخش منو انتخاب نمایید:

Transform→Random Number→Generators



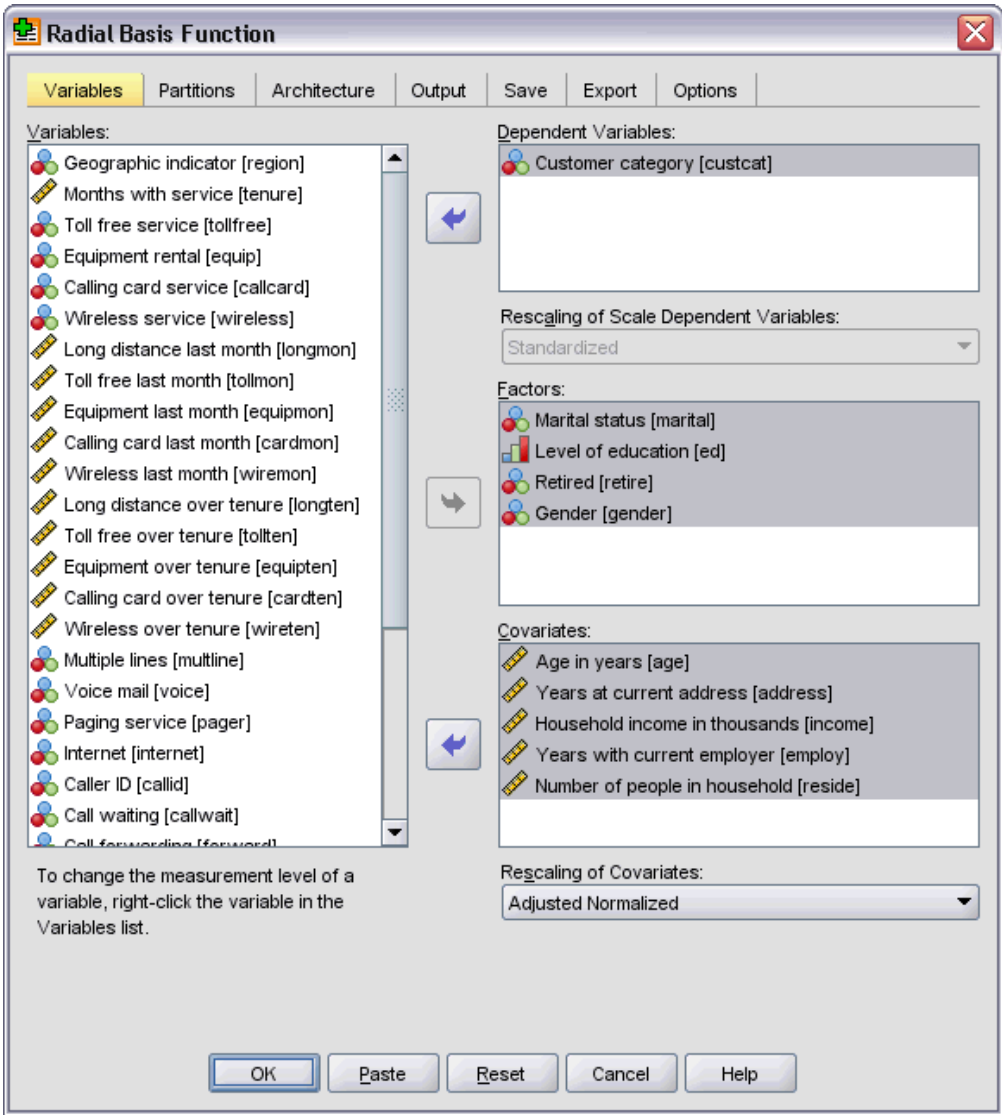

شکل ۳-۴۴

- گزینه set starting point را انتخاب نمایید.
- گزینه Fixed value را انتخاب کرده و عدد ۹۱۹۱۹۷۲ را به‌عنوان value در کادر تایپ کنید.
- برروی Ok، کلیک نمایید.

## راه‌اندازی آنالیزها

- جهت راه‌اندازی آنالیزها براساس تابع شعاع مدار، از بخش منو انتخاب نمایید.

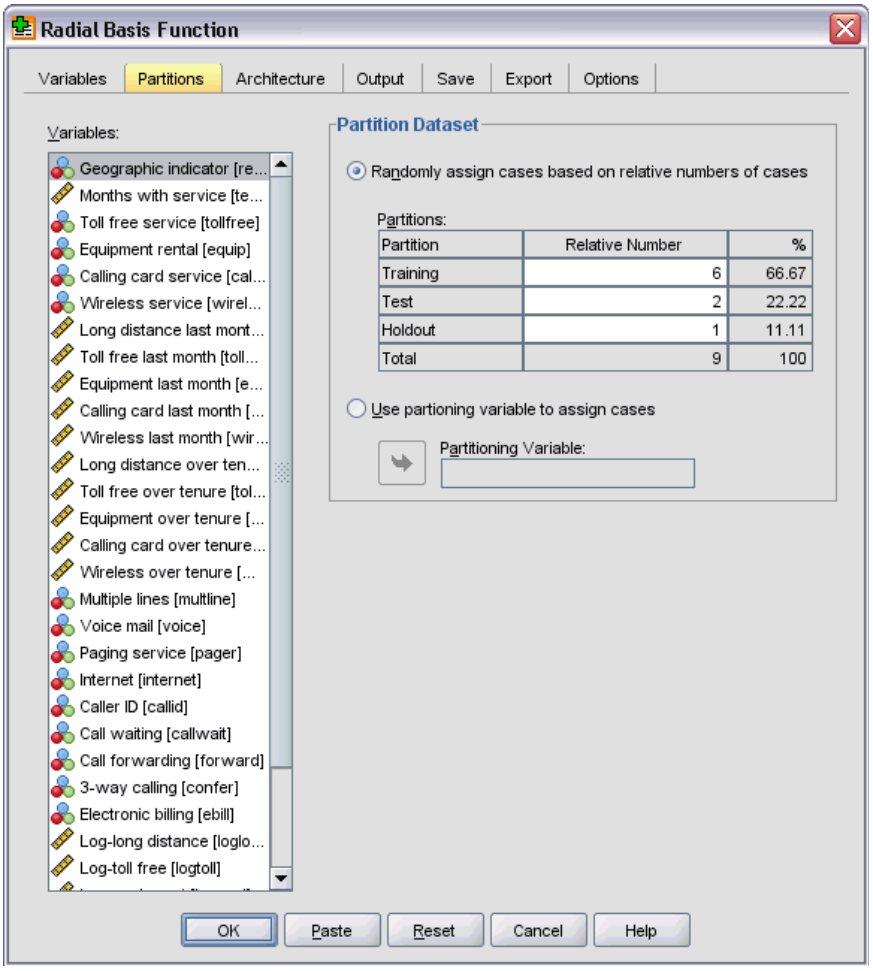
Analyze→Neural Networks→Radial Basis Function




شکل ۳-۴۵

- عبارت customer category را در قسمت dependent variable انتخاب کنید.
- عبارت‌های marital status ، Level of education، Retired و Gender را در قسمت Factors انتخاب نمایید.
- در قسمت متغیر کمکی عبارت‌های Age in years تا Number of people in household را انتخاب نمایید.

- روش Adjusted Normalized را در قسمت Rescalling of متغیر کمکی انتخاب نمایید.




*شکل ۳-۴۶*

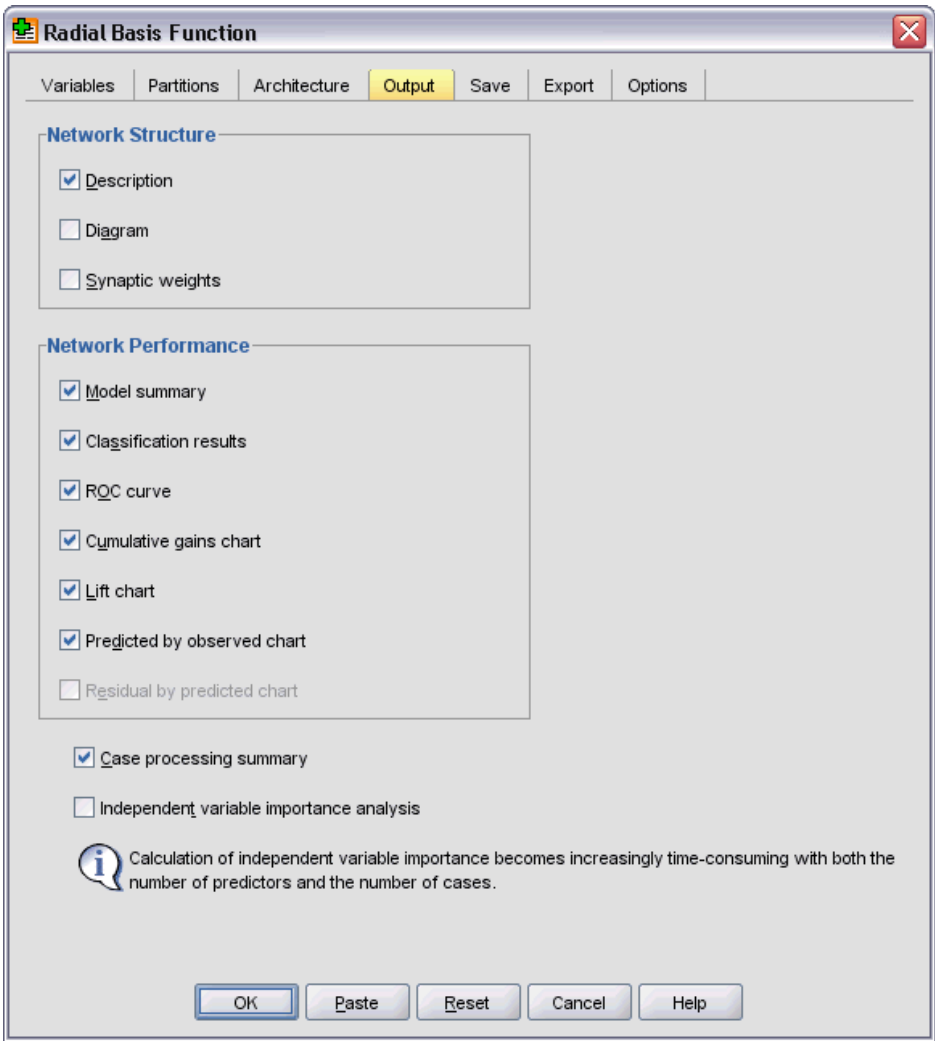
با مشخص نمودن تعداد تقریبی موارد، ایجاد تقسیم‌بندی‌های بسیار کوچک امکان‌پذیر و ساده است، اما ارائه آنان به‌صورت درصد می‌تواند مشکل باشد. بدین‌ترتیب که $\frac{۲}{۳}$ از داده‌ها را به‌عنوان نمونه آموزشی و $\frac{۲}{۳}$ از موارد باقی مانده را به نمونه‌های آزمایشی اختصاص می‌دهیم.

- عدد ۶ را در کادر مربوط به نمونه آموزشی تایپ نمایید.
- عدد ۲ را در کادر مربوط به نمونه آزمایشی تایپ نمایید.

- عدد ۱ را در کادر مربوط به نمونه جدا از هم نگه داشته شده تایپ نمایید.

تعداد کل موارد مشخص شده ۹ عدد می‌باشد، بدین‌ترتیب ۲۲/۲۲٪ از موارد را به نمونه‌های آزمایشی و $\frac{۱}{۹}$ آنها را که برابر با ۱۱/۱۱٪ از کل موارد می‌باشد، را به نمونه‌های جدا نگه داشته شده اختصاص می‌دهیم.

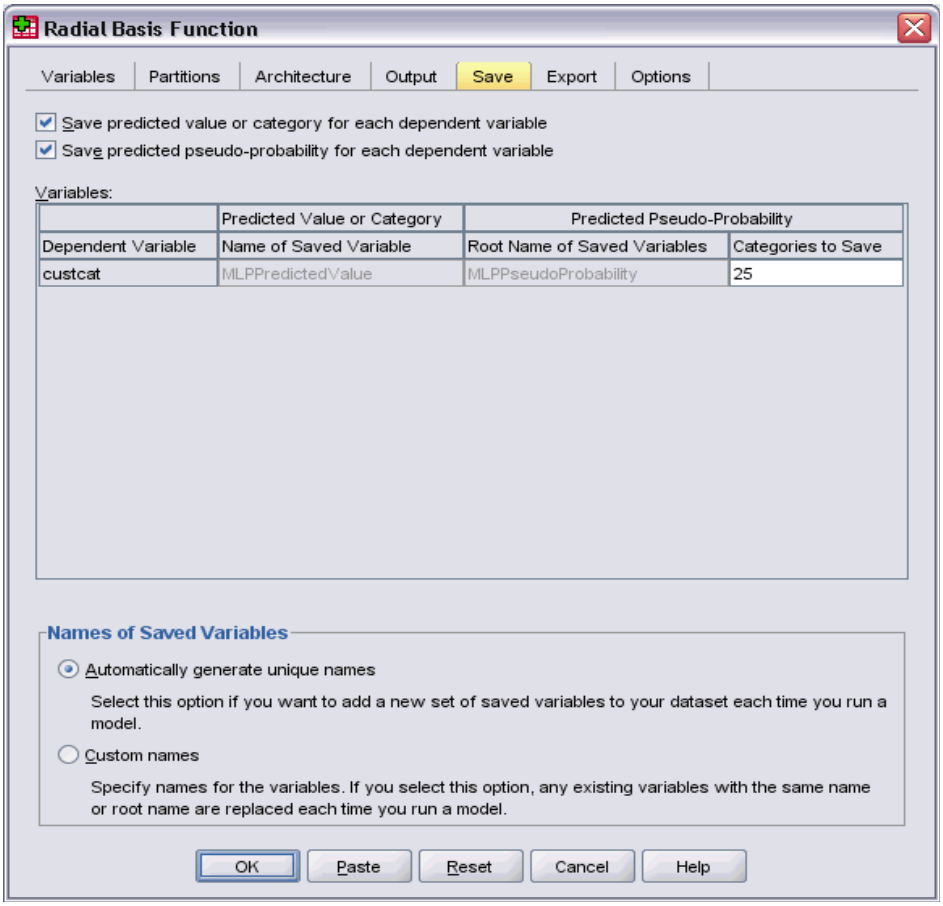
- برروی گزینه output کلیک نمایید.




*شکل ۳-۴۷*

- در بخش Network structure گزینه مربوط به Diagram را انتخاب نمایید.

- در بخش Network Performance گزینه‌های ROC curve، Cumulative gains chart، Lift chart و Predicted by observed chart را انتخاب کنید.
- برروی گزینه save کلیک کنید.




شکل ۳-۴۸

- دو گزینه ابتدایی را انتخاب کنید.
- برروی OK کلیک کنید.

## خلاصه فرایند انجام شده

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 665 | 66.5% |
| | Testing | 224 | 22.4% |
| | Holdout | 111 | 11.1% |
| Valid | | 1000 | 100.0% |
| Excluded | | 0 | |
| Total | | 1000 | |

شکل ۳-۴۹

این جدول نشان می‌دهد که ۶۶۵ مورد به نمونه آموزشی، ۲۲۴ مورد به نمونه آزمایشی و ۱۱۱ مورد به نمونه‌های جدا از هم نگه داشته شده اختصاص یافته است. هیچ یک از موارد از آنالیزها خارج نشده‌اند.

## اطلاعات شبکه

| | | | |
|---|---|---|---|
| Input Layer | Factors | 1 | Marital status |
| | | 2 | Level of education |
| | | 3 | Retired |
| | | 4 | Gender |
| | Covariates | 1 | Age in years |
| | | 2 | Years at current address |
| | | 3 | Household income in thousands |
| | | 4 | Years with current employer |
| | | 5 | Number of people in household |
| | Number of Units | | 16 |
| | Rescaling Method for Covariates | | Adjusted Normalized |
| Hidden Layer | Number of Units | | 9[a] |
| | Activation Function | | Softmax |
| Output Layer | Dependent Variables | 1 | Customer category |
| | Number of Units | | 4 |
| | Activation Function | | Identity |
| | Error Function | | Sum of Squares |

a.Determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

شکل ۳-۵۰

جدول اطلاعات شبکه، اطلاعات مربوط به شبکه عصبی را نمایش می‌دهد و برای حصول اطمینان از این که دسته‌بندی‌ها به درستی صورت گرفته‌اند، مفید می‌باشد. در این جا به موارد زیر توجه نمایید:

- تعداد واحدها در لایه ورودی برابر است با جمع مقدار کوواریانس‌ها و تعداد سطوح ضرایب.

  برای هر یک از دسته‌بندی‌های Marital status ، Level of education ، Retired و Gender یک واحد مجزا ایجاد گردیده است و هیچ یک از این دسته‌بندی‌ها همان‌طورکه در بسیاری از مدل‌های دیگر زائد در نظر گرفته می‌شوند، زائد در نظر گرفته نشده‌اند.
- به همین ترتیب، برای هر یک از دسته‌بندی‌های Customer category یک واحد خروجی مجزا ایجاد گردیده و در کل ۴ واحد در لایه خروجی، ساخته شده است.
- در کوواریانس‌ها، برای مقیاس‌بندی مجدد از روش Adjusted normalized استفاده شده است.
- فرایند مربوط به ساختار انتخاب خودکار، ۹ واحد را در لایه پنهان قرار داده است.
- ما بقی اطلاعات مربوط به فرایند به‌صورت پیش‌فرض از پیش در نظر گرفته شده‌اند.

## خلاصه مدل

| | | |
|---|---|---|
| Training | Sum of Squares Error | 235.969 |
| | Percent Incorrect Predictions | 61.8% |
| | Training Time | 2.72 |
| Testing | Sum of Squares Error | 80.851[a] |
| | Percent Incorrect Predictions | 62.9% |
| Holdout | Percent Incorrect Predictions | 59.5% |

Dependent Variable: Customer category

a. The number of hidden units is determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

*شکل ۳-۵۱*

این جدول اطلاعاتی را در مورد نتایج آموزش، آزمایش و کاربرد شبکه نهایی در مورد نمونه جدا از هم نگه داشته شده، نشان می‌دهد.

- مجموع مربعات خطا، از آنجایی که معمولاً در شبکه‌های RBF استفاده می‌شود، نمایش داده شده است. این تابع خطایی است که شبکه در طی فرایند آموزش و آزمایش، سعی دارد تا آن را به کمترین مقدار ممکن برساند.
- درصد پیش‌بینی‌های نادرست از جدول طبقه‌بندی گرفته شده است و در قسمت‌های بعدی بیشتر تشریح خواهد شد.

## طبقه‌بندی

| Sample | Observed | Predicted: Basic service | E-service | Plus service | Total service | Percent Correct |
|---|---|---|---|---|---|---|
| Training | Basic service | 64 | 0 | 66 | 45 | 36.6% |
| | E-service | 22 | 1 | 57 | 61 | .7% |
| | Plus service | 47 | 0 | 104 | 34 | 56.2% |
| | Total service | 29 | 1 | 49 | 85 | 51.8% |
| | Overall Percent | 24.4% | .3% | 41.5% | 33.8% | 38.2% |
| Testing | Basic service | 18 | 0 | 26 | 15 | 30.5% |
| | E-service | 15 | 0 | 16 | 22 | .0% |
| | Plus service | 11 | 0 | 39 | 15 | 60.0% |
| | Total service | 4 | 0 | 17 | 26 | 55.3% |
| | Overall Percent | 21.4% | .0% | 43.8% | 34.8% | 37.1% |
| Holdout | Basic service | 11 | 0 | 11 | 10 | 34.4% |
| | E-service | 4 | 0 | 9 | 10 | .0% |
| | Plus service | 10 | 0 | 19 | 2 | 61.3% |
| | Total service | 5 | 0 | 5 | 15 | 60.0% |
| | Overall Percent | 27.0% | .0% | 39.6% | 33.3% | 40.5% |

Dependent Variable: Customer category

*شکل ۳-۵۲*

این جدول نتایج خاص حاصل از استفاده از شبکه را نمایش می‌دهد. در هر مورد، پیش‌بینی انجام شده، دسته‌ای است که دارای بالاترین شبه احتمال پیش‌بینی شده باشد.

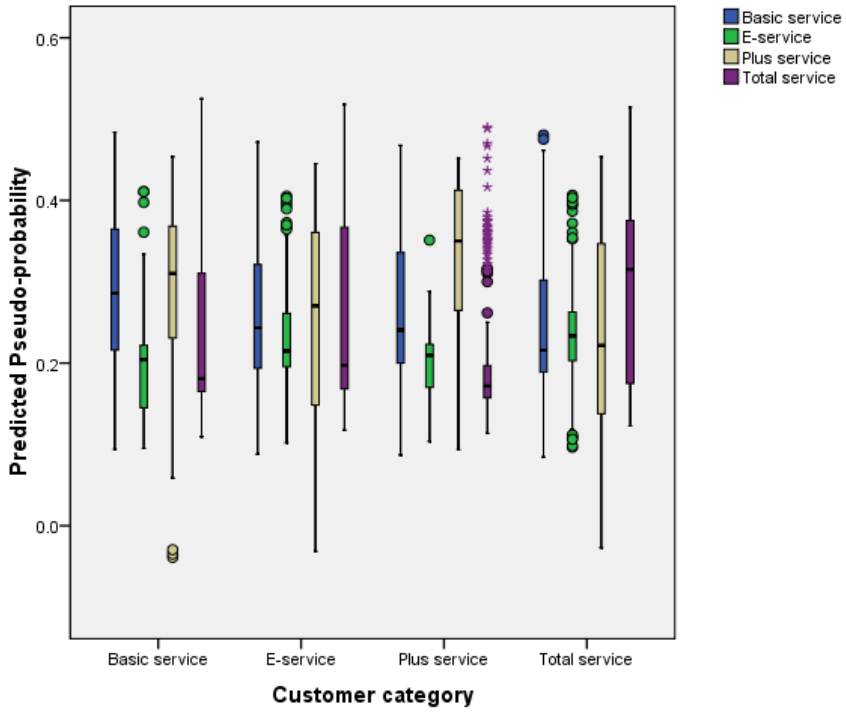
- سلول‌هایی که برروی خط اریب قرار می‌گیرند، پیش‌بینی‌های صحیح می‌باشند.
- سلول‌هایی که برروی خط اریب قرار نمی‌گیرند، پیش‌بینی‌های ناصحیح می‌باشند.

داده‌های مشاهده شده را ارائه دهید، آنگاه مدل تهی (مدلی است که فاقد پیش‌بینی‌کننده هاست) می‌تواند مشتریان را در گروه‌های چندگانه، طبقه‌بندی نماید، بنابراین مدل تهی در $\frac{۲۸۱}{۱۰۰۰} = ٪۲۸/۱$ از موارد درست عمل می‌نماید.

شبکه RBF ۱۰/۱٪ بیشتر از مدل تهی موفق عمل می‌نماید و یا به عبارت دیگر در ۳۸/۲٪ از موارد درست عمل می‌کند. به‌طور خاص مدل RBF در شناخت مشتریان خدمات اضافی (plus service) و مجموع کل خدمات (Total Service) بهتر عمل کرده و موفق است. با این وجود در مورد طبقه‌بندی مشتریان خدمات الکترونیک (E-service)، استثناً عملکرد ضعیفی را از خود نشان داده است. از این جهت ممکن است برای جدا نمودن این مشتریان نیاز به پیش‌بینی‌کننده دیگری داشته باشید. این مشتریان اغلب اوقات اشتباهاً در گروه‌های مشتریان خدمات اضافی (Plus service) و total service طبقه‌بندی می‌شوند. شرکت به سهولت می‌تواند مشتریان بالقوه خود را که معمولاً در گروه E-service قرار می‌گیرند، به خرید و استفاده بیشتر از خدمات ترغیب نماید.

طبقه‌بندی که بر پایه مواردی صورت می‌گیرد که از آن موارد جهت ساخت مدل استفاده شده است، بسیار خوش‌بینانه به نظر می‌رسد. به این معنا که نسبت طبقه‌بندی آنان متورم می‌شود. از نمونه جدا از هم نگه داشته شده جهت ارزیابی و اعتبارسنجی مدل استفاده می‌گردد، در اینجا، ۴۰/۲٪ از موارد توسط مدل به درستی طبقه‌بندی گردیده‌اند. این نمونه جدا از هم نگه داشته شده نسبتاً کوچک است و نشان می‌دهد که مدل شما در حقیقت ۲ بار از ۵ بار صحیح عمل می‌نماید.

## نمودار پیش‌بینی براساس مشاهده (Predicted by Observed Chart)




شکل ۳-۵۳

نمودار پیش‌بینی براساس مشاهده در مورد متغیرهای وابسته دارای دسته‌بندی، نشان‌دهنده مجموعه‌ای از نمودارهای میله‌ای است متشکل از شبه احتمال‌های پیش‌بینی شده، که در نتیجه ترکیبی از نمونه‌های آموزشی و آزمایشی به دست می‌آید.

نمودار میله‌ای که در منتهی‌الیه سمت چپ قرار گرفته است، نشان‌دهنده میزان شبه احتمال پیش‌بینی شده، دسته Basic Service و برای مواردی است که بنابر مشاهدات در این دسته قرار می‌گیرند.

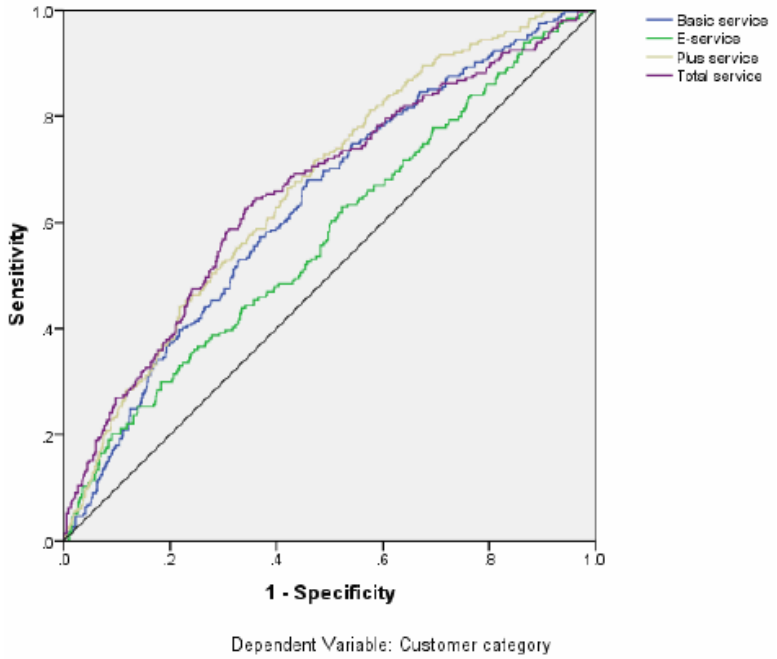
- نمودار میله‌ای بعدی که با حرکت به سمت راست بدان می‌رسیم، نشان‌دهنده میزان شبه احتمال پیش‌بینی شده دسته E-service، برای مواردی است که بنابر مشاهدات در دسته Basic service قرار گرفته‌اند.
- سومین نمودار میله‌ای، میزان شبه احتمال پیش‌بینی شده دسته Plus service را برای مواردی که در دسته Basic service قرار گرفته‌اند، را نشان می‌دهد. همان‌طورکه از جدول

طبقه‌بندی به خاطر می‌آورید. تعداد مشتریان متعلق به دسته Basic service که به‌صورت اشتباه در دسته Plus service قرار گرفته‌اند، به‌صورت تقریبی برابر با تعداد مشتریانی است که به درستی در این گروه دسته‌بندی شده‌اند. بنابراین، این نمودار میله‌ای تقریباً با نموداری که در منتهی‌الیه سمت چپ قرار گرفته است، متوازن است.

- چهارمین نمودار میله‌ای، میزان شبه احتمال پیش‌بینی شده دسته Total service را برای مواردی که در دسته Basic service قرار می‌گیرند، را نشان می‌دهد.

از آنجایی که تعداد متغیرهای هدف بیش از دو دسته می‌باشد، چهار نمودار میله‌ای اول نه تنها نسبت به خط افقی ۰/۵ بلکه نسبت به هیچ خطی متقارن نمی‌باشند. در نتیجه، تشریح این نمودار با بیش از ۲ متغیر هدف کار دشواری است، زیرا نمی‌توان با نگاه به بخشی از نمودار میله‌ای، نقطه متناظر آن مورد را در نمودار میله‌ای دیگری تشخیص داد.

## منحنی ROC (ROC curve)




شکل ۳-۵۴

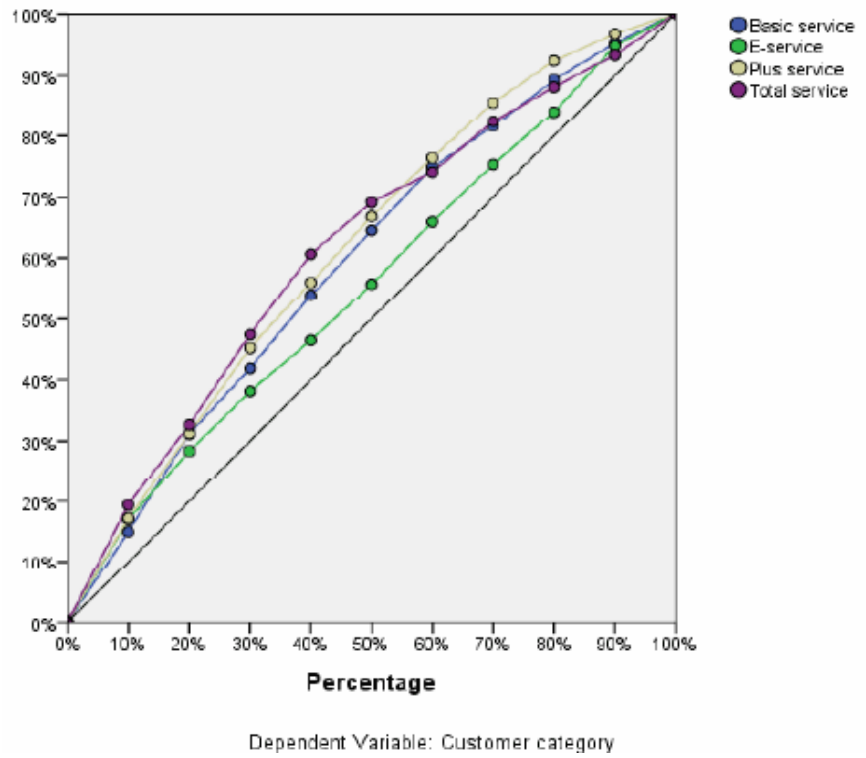
یک منحنی ROC، نمایشی بصری را از میزان حساسیت (sensitivity) در برابر specificity برای تمامی سطوح برش طبقه‌بندی نشان می‌دهد. نموداری که در اینجا نشان داده می‌شود دارای ۴ منحنی بوده که هر یک متعلق به یکی از متغییرهای هدف است.

توجه داشته باشید که این نمودار براساس مجموعه‌ای ادغامی از نمونه‌های آزمایشی و آموزشی ترسیم می‌شود. به‌منظور ترسیم نمودار ROC برای نمونه جدا از هم نگهداری شده، فایل موردنظر را توسط متغییر تفکیکی (partition variable) به ۲ بخش تقسیم نمایید و فرآیند مربوط به منحنی ROC را با شبه احتمالات پیش‌بینی شده فعال نمایید.

| | | Area |
|---|---|---|
| Customer category | Basic service | .635 |
| | E-service | .573 |
| | Plus service | .668 |
| | Total service | .659 |

سطح زیر منحنی (area under curve) خلاصه‌ای عددی را از منحنی ROC ارائه می‌دهد. مقادیر موجود در این جدول، برای هر یک از دسته‌بندی‌ها، احتمال این که شبه احتمال پیش‌بینی شده آن دسته برای مواردی که به‌صورت تصادفی در آن انتخاب شده‌اند بیشتر از مقادیر شبه احتمال پیش‌بینی شده‌ای است که با انتخاب تصادفی خارج از دسته به‌دست می‌آیند، را نشان می‌دهد. برای مثال، با انتخاب تصادفی مشتریان در دسته مشتریان خدمات اضافی(plus service) و انتخاب تصادفی که در دسته‌های مشتریان خدمات پایه (basic service)، مشتریان خدمات الکترونیک (E-service) و یا مجموع کل خدمات (Total service) انجام می‌گیرد، احتمال این که شبه احتمال پیش‌بینی شده توسط مدل در حالت پیش‌فرض برای مشتری موجود در دسته خدمات اضافی (plus service) بیشتر باشد، برابر با ۰/۶۶۸ است.

## Cumulative gains and lift charts

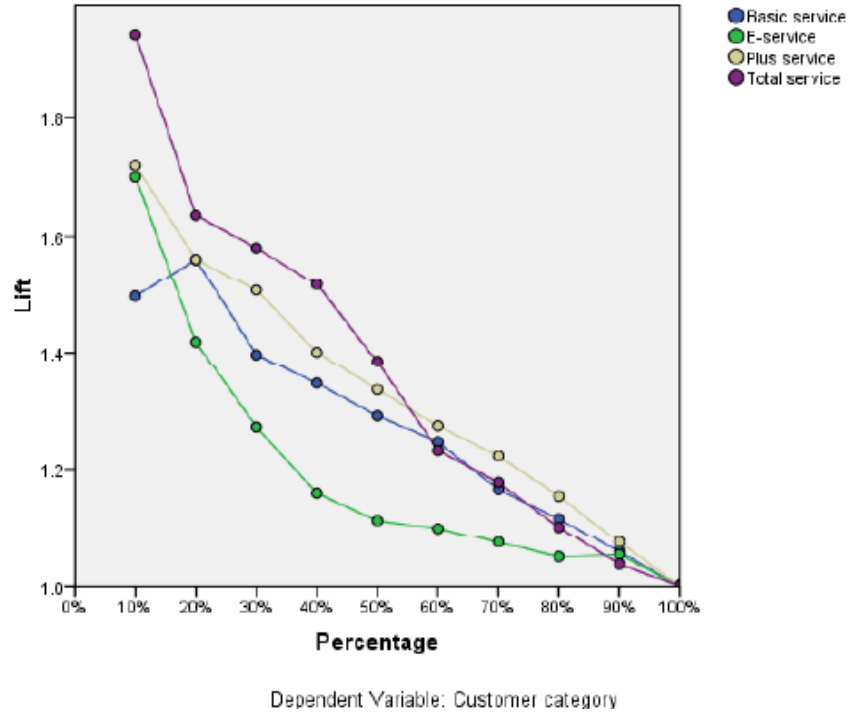



شکل ۳-۵۵

این نمودار، درصدی از تعداد مواردی را که در یک دسته مشخص به‌وسیله هدف قرار دادن درصد معینی از کل موارد در دسترس، قرار می‌گیرند، را نشان می‌دهد. برای مثال، اولین نقطه‌ای که در نمودار برای دسته مجموع کل خدمات (Total service) وجود دارد حدوداً در منطقه (۲۰٪ ،۱۰٪) قرار گرفته است و این بدان معناست که چنانچه شما مجموعه داده‌ای را توسط شبکه ثبت نمایید و سپس تمامی آن داده‌ها را با توجه به میزان شبه احتمال پیش‌بینی شده مجموع کل خدمات (Total service) مرتب نمایید، آنگاه می‌توانید انتظار داشته باشید که در میان ۱۰٪ اول این داده‌ها، ۲۰٪ از کل مواردی که در واقع می‌باید در دسته مجموع کل خدمات (Total service) قرار گیرند، جای می‌گیرند. به همین ترتیب، در میان ۲۰٪ اول داده‌های مرتب شده در حدود ۳۰٪ از کل مواردی که باید در این دسته قرار گیرند، قرار می‌گیرند، و...

چنانچه ۱۰۰٪ داده‌های ثبت شده را انتخاب نمایید، تمامی موارد موردنظر خود را در آن دسته مشخص به‌دست خواهید آورد.

خط اریبی را که مشاهده می‌نمایید، یک منحنی مبنا است که در آن، چنانچه شما ۱۰٪ از داده‌های ثبت شده را انتخاب نمایید، می‌توانید انتظار داشته باشید که حدوداً ۱۰٪ از تمامی مواردی را که در واقع در هر دسته‌ای قرار می‌گیرد را شامل شود. به هر میزان که از این منحنی مبنا بالاتر قرار گیریم، تعداد مواردی که در دسته موردنظر قرار می‌گیرد، بیشتر می‌شود.




*شکل ۳-۵۶*

نمودار lift chart از نمودار cumulative gain chart (نمودار بهره تجمعی) به‌دست می‌آید. مقادیر موجود برروی محور y متناظر با نسبت بهره تجمعی برای هر یک از منحنی‌ها تا خط مبنا است. بنابراین میزان lift، برای دسته Total service در نقطه ۱۰٪ در حدود $\frac{٪۲۰}{٪۱۰}=۲$ می‌باشد. این نسبت نقطه‌نظر جدیدی را برای نگاه کردن به اطلاعات موجود در نمودار بهره تجمعی، فراهم می‌سازد.

**توجه:** نمودارهای cumulative gains and lift charts براساس مجموعه ترکیبی از نمونه‌های آزمایشی و آموزشی ترسیم می‌شوند.

# پیوست

# فایل‌های نمونه

فایل‌های نمونه همراه با نرم‌افزار برروی سیستم نصب گردیده است و می‌توان آن را در فهرست نمونه‌ها مشاهده نمود. این نمونه‌ها به زبان‌های انگلیسی، فرانسه، آلمانی، ایتالیایی، ژاپنی، کره‌ای، لهستانی، روسی، چینی ساده شده، اسپانیایی و چینی سنتی، آماده و در دسترس می‌باشند.

تمام فایل‌های نمونه را به تمامی زبان‌های عنوان شده نمی‌توان یافت، چنانچه فایل نمونه‌ای به زبان موردنظر شما وجود نداشت آنگاه می‌بایست از نمونه انگلیسی آن استفاده نمایید.

**توضیح:**

در ادامه تشریح مختصری از فایل‌های نمونه که در مثال‌های متعددی استفاده می‌گردند، آورده شده است.

- **Accidents.sav** . فایلی است که شامل داده‌های فرضی در مورد مطالعات انجام شده توسط یک شرکت خدمات بیمه، پیرامون تأثیر مؤلفه‌های سن و جنس افراد برروی میزان تصادفات رانندگی انجام شده در یک ناحیه خاص.
- **Adl.sav** . فایلی است شامل داده‌های فرضی که تلاش بر اندازه‌گیری میزان سودمندی انواع مختلف درمان بیمارانی که دچار شکستگی شده‌اند، را دارد. متخصصین فیزیوتراپی، به‌صورت تصادفی بیماران دچار شکستگی را در یکی از دو گروه زیر قرار می‌دهند. گروه اول درمان‌های فیزیوتراپی استاندارد را دریافت کرده و گروه دوم خدمات احساسی دیگری را علاوه بر فیزیوتراپی استاندارد، دریافت می‌نمایند. با گذشت سه ماه از طول مدت درمان، توانایی هر یک از بیماران در انجام امور روزمره زندگی‌شان مورد بررسی قرار می‌گیرد و به‌عنوان یک متغیر ترتیبی ثبت می‌گردد.
- **Advert.sav** . فایلی است شامل داده‌های فرضی که در آن تلاش‌های انجام گرفته از سوی یک خرده فروش برای بررسی رابطه میان میزان فروش صورت گرفته و هزینه انجام شده برروی امر تبلیغات بررسی شده است. بدین منظور، سوابق فروشی که در گذشته صورت گرفته است را به همراه میزان هزینه تبلیغات مربوط به آن را انتخاب کرده‌ایم.
- **Aflatoxin.sav** . فایلی است شامل داده‌های فرضی که در آن به بررسی تأثیر سم aflatoxin برروی محصول گندم می‌پردازیم. سمی که میزان غلط آن در دامنه گسترده‌ای متنوع بوده و برروی بازده محصول تأثیر می‌گذارد. یک پردازشگر grain، ۱۶ نمونه مختلف را از هر

یک از ۸ انبار محصول دریافت نموده و میزان سم alfatoxin را در مقیاس PPB (تعداد واحد در میلیارد) مورد سنجش قرار می‌دهد.

- **Aflatoxin 20 .sav** . این فایل شامل داده‌های مربوط به اندازه‌گیری میزان سم aflatoxin موجود در ۱۶ نمونه مربوط به انبارهای محصول ۴ و ۸ موجود در فایل aflatoxin.sav است.
- **Anorectic.sav** . در طی تلاش‌های انجام گرفته برای استاندارد نمودن شیوه شناسایی بیماری‌های کم خوری و پرخوری، محققان ۵۵ مورد از بیماران بالغی که دچار مشکلات تغذیه‌ای بودند را مورد مطالعه قرار داده‌اند. هر یک از بیماران در طی ۴ سال، ۴ بار مورد بررسی قرار گرفته که در کل ۲۲۰ مورد بررسی را تشکیل می‌دهد. در هر بار مشاهده وجود و یا عدم وجود ۱۶ نشانه بیماری در هر یک از بیماران بررسی می‌گردد. در میان نشانه‌های ثبت شده، موارد زیر آورده نشده است. بیمار شماره ۷۱ در دومین بار بررسی، بیمار شماره ۷۶ در اولین بررسی و بیمار شماره ۴۷ در سومین بررسی.
- **Autoaccidents.sav** . این فایل دربرگیرنده اطلاعات فرضی پیرامون بررسی‌های یک شرکت بیمه جهت مدل‌سازی تعداد تصادفات اتومبیل و ارتباط آن با تعداد رانندگان و همچنین سن و جنس هر کدام از این افراد است. هر یک از موارد، مختص به راننده‌ای خاص بوده که اطلاعاتی در مورد جنسیت و سن راننده به همراه تعداد تصادفات روی داده توسط آن شخص در طی ۵ سال گذشته در آن ثبت گردیده است.
- **Band.sav** . این فایل حاوی اطاعات فرضی مربوط به میزان فروش هفتگی آلبوم‌های یک گروه موسیقی است. داده‌های مربوط به ۳ متغیر پیش‌بینی نیز در این فایل وجود دارد.
- **Bankloan.sav** . این فایل شامل داده‌هایی فرضی در مورد تلاش‌های یک بانک جهت کاهش تعداد مشتریان بدحساب خود است. در این فایل اطلاعات مالی و آماری مربوط به ۸۵۰ مشتری سابق و مشتری بالقوه بانک آورده شده است. ۷۰۰ مورد اول مشتریانی هستند که سابقاً از بانک وام دریافت نموده‌اند و ۱۵۰ مورد آخر نیز مشتریان بالقوه‌ای هستند که بانک برای دسته‌بندی آنها به مشتریانی با ریسک اعتباری زیاد یا کم به آنان نیاز دارد.
- **Bankloan-binning.sav** . این فایل شامل اطلاعات مالی و آماری فرضی در رابطه با ۵۰۰۰ مشتری قدیمی بانک است.

- **Behavior.sav** . در آزمونی کلاسیک، از ۵۲ دانش آموز خواسته شد تا ترکیبی از ۱۵ موقعیت و ۱۵ عکس‌العمل مطابق با آنان را امتیازدهی کنند، بدین صورت که امتیازها در مقیاس از ۱۰ تا ۰ بوده و عدد صفر به نامناسبترین و عدد ۱۰ به مناسبترین عکس‌العمل در برابر هر یک از موقعیت‌ها تعلق می‌گیرد. سپس از اعداد به‌دست آمده برای هر شخص میانگین گرفته می‌شود.
- **Behavior-ini.sav** . این فایل شامل داده‌های اولیه مربوط به راه‌حلی دو بعدی برای فایل behavior.sav است.
- **Brakes.sav** . این فایل شامل داده‌های فرضی در مورد فرایند کنترل کیفیت کارخانه‌ای است که به تولید لنت‌های ترمز اتومبیل‌های پیشرفته می‌پردازد. فایل محتوی داده‌های مربوط به اندازه‌گیری قطر ۱۶ نمونه از هر یک ۸ دستگاه تولید لنت می‌باشد. اندازه موجود در نقشه و تعیین شده در مرحله طراحی برای هر یک از لنت‌ها برابر با ۳۲۲ میلی‌متر است.
- **Breakfast.sav** . در یک مطالعه کلاسیک، از ۲۱ دانشجوی رشته MBA که در مدرسه Wharton مشغول به تحصیل‌اند خواسته شده تا ۱۵ مورد از مواردی که در وعده صبحانه مصرف می‌شود را به ترتیب رتبه‌بندی نمایند، بدین‌ترتیب که عدد ۱ به موردی اختصاص داده می‌شود که بیشترین علاقه بدان وجود دارد و عدد ۱۵ به موردی که کمترین علاقه به مصرف آن در وعده صبحانه در شخص وجود دارد. میزان علاقه افراد تحت شش دسته‌بندی از ترجیح کلی تا یک وعده غذایی مختصر تنها شامل نوشیدنی قرار می‌گیرد.
- **Breakfast-overall.sav** . داده‌های موجود در این فایل تنها حاوی داده‌های مربوط به عبارت Overall preference و میزان علایق مربوط به موارد موجود در وعده صبحانه که در این دسته قرار می‌گیرند، می‌باشد.
- **Broadband-1.sav** . این فایل شامل داده‌های فرضی مربوط به تعداد مشترکان یک شبکه خدمات‌دهی گسترده می‌باشد که با توجه به منطقه جغرافیایی ثبت شده‌اند، این داده‌ها شامل تعداد مشترکان در ۸۵ نقطه جغرافیایی در طی یک دوره ۴ ساله است.
- **Broadband-2.sav** . این فایل نیز مشابه فایل broadband-1.sav است، با این تفاوت که در این فایل داده‌های مربوط به سه ماه بیشتر از فایل ۱ آورده شده است.
- **Car-insurance-claims.sav** . یک مجموعه داده که بیش از این توسط (McCullagh and Nelder,1998) آنالیز و ارائه گردیده و در مورد خسارت‌های وارد به

ماشین‌ها می‌باشد. مقدار میانگین این خسارت‌ها را می‌توان به‌گونه مدل نمود که از توزیع گاما پیروی نمایند؛ این کار را می‌توان با استفاده از یک تابع ارتباطی معکوس جهت برقراری ارتباط میان میانگین متغیر وابسته با یک مجموعه ترکیبی خطی از طول مدتی که شخص بیمه بوده، نوع خودرو و عمرخودرو انجام داد. از تعداد خسارات فایل شده می‌توان برای مقیاس وزن‌دهی استفاده نمود.

- **Car-sales.sav** . این فایل محتوی اطلاعات فرضی پیرامون فروش، لیست قیمت‌ها و ویژگی‌های فیزیکی مدل‌های مختلف و متنوع خودرو است. اطلاعات مربوط به لیست قیمت‌ها و ویژگی‌های فیزیکی به ترتیب از سایت Edmunds.com و سایت‌های مربوط به تولیدکنندگان کسب گردیده‌اند.
- **Carpet.sav** . در یک مثال بسیار معروف (Green and Wind,1973) آمده است، شرکتی که قصد ساخت یک پاک کننده فرش جدید را دارد، علاقه‌مند است تأثیر ۵ عامل زیر را برروی میزان علاقه مشتریان به استفاده از این محصولات برررسی نماید. طرح بسته‌بندی محصول، عنوان تجاری محصول، قیمت، داشتن تصدیق‌نامه معتبر و داشتن گارانتی که به‌وسیله آن پول مشتری پس داده شود. در زمینه طرح بسته‌بندی محصول سه سطح وجود دارد که وجه تمایز آنان محل قرارگیری برس همراه با محصول است. سه عنوان تجاری (K2R,Glory,Bissell)، سه سطح قیمت و برای هر یک از دو عامل آخر دو سطح برای محصول در نظر گرفته شده است. مشتریان براساس این عوامل ۲۲ پروفایل را دسته‌بندی می‌نمایند. متغیر Preference شامل رتبه‌بندی میان متوسط رتبه‌هایی است که به هر یک از پروفایل‌ها داده شده است. هر چه عدد مربوط به رتبه داده شده کمتر باشد نشان‌دهنده میزان علاقه بیشتر مشتریان است. این متغیر سنجش کلی را برای هر یک از پروفایل‌ها ارائه می‌دهد.
- **Carpet-prefs.sav** . این فایل نیز بر پایه مثالی که در فایل carpet.sav از آن استفاده شده بود، قرار دارد، با این تفاوت که این فایل محتوی رتبه‌بندی حقیقی است که از میان هر ۱۰ مشتری انتخاب گردیده است. از مشتریان خواسته شده است که از میان ۲۲ پروفایل تولیدی رتبه‌بندی را انجام دهند. متغیرهای PREF1 تا PREF22 هر کدام معرف یکی از این پروفایل‌ها است.

- **Catalog.sav** . این فایل شامل اطلاعات فرضی در مورد میزان فروش سه محصول یک شرکت تولیدی است. در این فایل، همچنین داده‌های مربوط به پنج متغیر پیش‌بینی محتمل نیز آورده شده است.
- **Catalog-seasfac.sav** . این فایل نیز شبیه فایل Catalog.sav است با این تفاوت که در آن مجموعه‌ای از محاسبات مربوط به عوامل خطی نیز آورده شده است.
- **Cellular.sav** . این فایل شامل داده‌هایی فرضی در مورد تلاش‌های انجام شده از سوی یک شرکت تولید تلفن سلولی است، برای کاهش میزان تکان‌های شدید. در این نوع سلول‌ها، میل ذاتی انجام این تکان‌های شدید وجود دارند و برای این که مورد محاسبه قرار گیرد در دامنه‌ای از ۰ تا ۱۰۰ ثبت می‌گردند. مواردی که از ۵۰ به بالا ثبت گردند، مواردی هستند که باعث ایجاد تغییرات می‌شوند.
- **Ceramics.sav** . این فایل شامل داده‌هایی فرضی است پیرامون تلاش‌های انجام گرفته یک تولیدکننده جهت محاسبه این که آیا آلیاژ با ترکیب جدید می‌تواند مقاومت حرارتی بیشتری نسبت به آلیاژهای استاندارد داشته باشد یا خیر. هر یک از موارد موجود در این فایل نشان‌دهنده آزمایش مجزای هر یک از آلیاژها است (دمایی که در آن مقاومت حرارتی از بین می‌رود ثبت شده است)
- **Cereal.sav** . این فایل شامل اطلاعاتی فرضی در مورد بررسی انجام شده در میان ۸۸۰ نفر می‌باشد که درباره موارد مورد علایق‌شان برای مصرف در وعده صبحانه پرسش شده است و همچنین مواردی همچون سن، جنس، وضعیت تأهل و این که آیا هر یک از این افراد دارای زندگی فعالی می‌باشند یا خیر (براین اساس که در هفته حداقل ۲ بار ورزش نمایند) در این فایل آورده شده است.
- **Clothing-defects.sav** . این فایل شامل اطلاعاتی فرضی پیرامون فرایند کنترل کیفیت در یک کارخانه تولید پوشاک است. بازرسان از هر سری از لباس‌های تولید شده نمونه‌هایی را برداشته و تعداد لباس‌های غیرقابل قبول را شمارش می‌نمایند.
- **Cantacts.sav** . این فایل حاوی داده‌هایی فرضی در مورد لیست‌های ارتباط میان دسته‌ای از نمایندگی‌های فروش کامپیوتر که با یکدیگر همکاری می‌نمایند است. هر یک از تماس‌ها با توجه به محل ساختمان شرکت که در آن مشغول به فعالیت هستند و همچنین مرتبه شرکت طبقه‌بندی می‌شوند. در این فایل موارد دیگری همچون میزان فروشی که شرکت در آخرین

بار فروش داشته، زمانی که از این فروش گذشته و ابعاد و میزان ارتباطات شرکت نیز ثبت شده است.

- **Creditpromo.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های انجام شده از سوی یک فروشگاه بزرگ برای بررسی تأثیر تبلیغات انجام شده برروی کارت‌های اعتباری است. بدین منظور، ۵۰۰ نفر از دارندگان این کارت‌ها به‌صورت تصادفی انتخاب شدند.
- **Customer-dbase.sav** . این فایل شامل داده‌های فرضی درباره تلاش‌های یک شرکت جهت استفاده از اطلاعات در دسترس خود، برای ارائه پیشنهادهای ویژه به مشتریانی است که به احتمال بیشتری این پیشنهادها را می‌پذیرند. یک دسته فرعی از مشتریان که به‌صورت تصادفی انتخاب شده‌اند را در نظر گرفته و به آنان پیشنهادهای ویژه‌ای داده شده است، سپس عکس‌العمل آنان را مشاهده نموده و ثبت گردیده است.
- **Customer-information.sav** . این فایل شامل، داده‌های فرضی است که پست الکترونیک مشتریان و اطلاعاتی همچون نام و آدرس مشتریان در آن وجود دارد.
- **Customers-model.sav** . این فایل شامل داده‌های فرضی در مورد افرادی است که توسط یک گروه بازاریاب مدنظر قرار می‌گیرند. این داده‌ها، شامل اطلاعات مربوط به آمارگیری نفوس، تاریخچه‌ای از میزان خرید و این که آیا هر یک از این افراد به گروه بازاریاب پاسخ می‌دهد یا خیر، می‌باشند. هر مورد نماینده یک شخص خاص است.
- **Customers-new.sav** . این فایل شامل داده‌های فرضی در مورد افرادی است که پتانسیل آن که در یک گروه بازاریاب قرار گیرند، را دارا هستند. این داده‌ها شامل اطلاعات مربوط به آمارگیری نفوس و خلاصه‌ای از تاریخچه میزان خرید برای هر یک از افراد می‌باشد. هر مورد نماینده یک شخص خاص است.
- **Debate.sav** . این فایل شامل داده‌هایی فرضی است که در آن عکس‌العمل‌های عده‌ای از بینندگان یک مناظره سیاسی را قبل و بعد از مناظره ثبت می‌نماید.
- **Debate-aggregate.sav** . این فایل شامل داده‌های فرضی است که عکس‌العمل‌های موجود در فایل debate.sav را جمع‌بندی می‌نماید. هر یک از موارد به یک طبقه‌بندی از میزان علاقه‌مندی قبل و بعد از مناظره، واکنش نشان می‌دهند.

- **Demo.sav** . این فایل شامل داده‌های فرضی است از یک پایگاه داده که در آن اسامی مشتریان که خرید نموده‌اند قرار گرفته است و این ثبت اسامی برای فرستادن ماهیانه پیشنهاد خرید به آنان، صورت می‌پذیرد. این که مشتری به پیشنهاد پاسخ می‌دهد یا نمی‌دهد نیز ثبت می‌گردد.
- **Demo-cs-1.sav** . این فایل شامل داده‌های فرضی در مورد اولین فرم یک شرکت برای جمع‌آوری اطلاعات مربوط به ممیزی و قرار دادن آنها در قالب یک پایگاه داده است. هر مورد به یک شهر و منطقه جغرافیایی، استان و ناحیه مختلف تعلق داشته که این اطلاعات نیز در فایل ثبت شده است.
- **Demo-cs-2.sav** . این فایل شامل داده‌های فرضی در مورد فرم دوم یک شرکت برای جمع‌آوری و تألیف پایگاه داده خود برای اطلاعات مربوط به ممیزی‌ها است، هر مورد به یک واحد خانوار مختلف از شهرهایی است که در قدم اول انتخاب شده‌اند و منطقه جغرافیایی، استانی، ناحیه، شهر، بخش فرعی و مشخصات واحد نیز ثبت گردیده‌اند. اطلاعات مربوط به نمونه‌های گرفته شده در دو مرحله اول طراحی نیز در فایل موجود می‌باشد.
- **Demo-cs.sav** . این فایل شامل داده‌های فرضی است که در آن برای اطلاعات ممیزی انتخاب شده از یک طرح نمونه‌گیری پیچیده استفاده می‌شود. هر مورد متعلق به یک واحد خانوار مختلف می‌باشد و اطلاعات مربوط به نمونه‌گیری و تنوع موجود در آمارگیری نفوس نواحی مختلف نیز در این فایل ثبت شده‌اند.
- **Dietstudy.sav** . این فایل شامل داده‌های فرضی در مورد نتایج غذایی "Stillman diet" است. هر مورد به یک شخص تعلق داشته و وزن و میزان تری‌گلیسیرید خون، قبل و بعد از استفاده از رژیم ثبت گردیده است.
- **Dischargedata.sav** . این فایل شامل داده‌های فرضی در مورد نحوه و الگوی استفاده فصلی از مرکز بیمارستانی Manitoba در جهت سیاست‌گذاری‌های مربوط به سلامت است.
- **Dvdplayer.sav** . این فایل شامل داده‌های فرضی در مورد پیشرفت دستگاه‌های جدید بخش DVD است. تیم بازاریابی با استفاده از یک نمونه اولیه، دسته داده متمرکز شده‌ای را انتخاب کرده‌اند. هر مورد مربوط به یک مصرف‌کننده مشخص بوده که همراه با اطلاعات

مربوط به آن، اطلاعاتی از قبیل آمارگیری نفوس و پاسخ ارائه شده توسط آنان به سؤالاتی در مورد نمونه اولیه نیز در این فایل ثبت شده است.

- **Flying.sav** . این فایل حاوی فاصله هوایی میان ۱۰ شهر ایالات متحده امریکا برحسب واحد مایل است.
- **German-credit.sav**. داده‌های موجود در این فایل از مجموعه داده‌ای German Credit در پایگاه داده‌ای موجود در دانشگاه کالیفرنیا به‌دست آمده است.
- **Grocery-1month.sav** . این فایل شامل داده‌های فرضی در فایل Grocery-coupons.sav است که اطاعات مربوط به میزان خرید هفتگی آن پنهان گردیده است. بنابراین هر مورد به اطلاعات مربوط به مشتری خاص تعلق دارد. برخی از متغیرها در نتیجه این که به‌صورت هفتگی تغییر می‌نمایند، ناپدید می‌شوند و میزان پول پرداختی ثبت شده در این فایل مجموع، پول پرداختی در طول ۴ هفته‌ای است که مطالعه صورت می‌پذیرد.
- **Grocery-coupons.sav** . این فایل شامل داده‌های فرضی است که توسط یک متصدی چرخه خواروبار که علاقه‌مند است خرید معمول مشتریان خود را ثبت و بررسی نماید، فراهم آمده است. هر یک از مشتریان برای مدت ۴ هفته زیر نظر قرار گرفته می‌شود و هر یک از موارد حاوی اطلاعات مربوط به خرید و چگونگی خرید هر یک از مشتریان به‌صورت هفتگی است. اطلاعات از قبیل این که مشتریان از چه مکان‌هایی خرید می‌نمایند و یا به چه میزان در هفته برای تهیه خواروبار هزینه می‌نمایند نیز در فایل ثبت گردیده است.
- **Guttma.sav** . Bell در سال ۱۹۶۱، جدولی را جهت نشان دادن گروه‌های اجتماعی تهیه و ارائه نمود. Guttman در سال ۱۹۶۸، از بخشی از این جدول که در آن پنج متغییر مسائلی همچون روابط اجتماعی، احساس تعلق به یک دسته، نزدیکی و مشابهت فیزیکی اعضای یک گروه را توصیف می‌نمایند، استفاده کرد. در این قسمت از جدول شکل و ساختار روابط توسط هفت دسته اجتماعی تقسیم‌بندی شده شکل می‌گیرد که عبارتند از جمعیت زیاد (برای مثال، افرادی که برای یک بازی فوتبال جمع می‌شوند) مخاطبین (برای مثال، افرادی که به دیدن یک نمایش تئاتر رفته و یا به یک سخنرانی گوش می‌دهند) عموم (برای مثال، مخاطبین روزنامه و یا تلویزیون) گروه‌های اولیه (خودمانی) گروه‌های ثانویه (داوطلب) و جوامع مدرن (از دست دادن هم پیمان‌های جمعی که در نتیجه مشابهت‌های فیزیکی زیاد ایجاد شده و نیاز به خدمات ویژه‌ای نیز وجود دارد.)

- **HealthPlans.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های یک گروه بیمه برای ارزیابی عملکرد برنامه درمانی مختلف در گروه‌های کوچکی از کارفرماها است. از ۱۲ نفر از این کارفرماها خواسته شد تا برنامه‌ها را با توجه به این که چه میزان علاقه‌مند هستند که برای کارمندان خود هزینه نمایند، رتبه‌بندی کنند. هر مورد متعلق به یک کارفرمای مجزا و عکس‌العمل آن در مقابل هر یک از برنامه‌ها است.
- **Health-Funding.sav** . این فایل شامل داده‌های فرضی در مورد میزان سرمایه‌گذاری در بخش سلامت (مبلغ به ازای هر ۱۰۰ نفر جمعیت)، میزان بیماری (میزان بیماری به ازای هر ۱۰۰۰۰ نفر جمعیت) و ملاقات با تأمین‌کنندگان سلامتی (میزان ملاقات به ازای هر ۱۰۰۰۰ نفر جمعیت) است. هر مورد معرف اطلاعات مربوط به شهرهای مختلف می‌باشد.
- **Hivassay.sav** . این فایل شامل داده‌های فرضی در مورد یک آزمایشگاه دارویی است که تلاش دارد روشی را برای تشخیص سریع آلودگی HIV ایجاد نماید. این آزمایشگاه برروی ۲۰۰۰ نمونه‌خونی که نیمی از آنها به HIV آلوده بوده و نیمی دیگر بدون آلودگی بودند بررسی‌های خود را انجام داده است.
- **Hourlywagedata.sav** . این فایل شامل داده‌های فرضی در مورد دستمزد ساعتی پرستاران است که در مطب پزشک و یا بیمارستان مشغول به کار هستند و دارای سطوح مختلفی از لحاظ تجربه کاری می‌باشند.
- **Insure.sav** . این فایل شامل داده‌های فرضی در مورد یک شرکت خدمات بیمه است که در زمینه عوامل خطر مطالعه می‌نماید. این عوامل خطر مشخص می‌نماید که آیا یک متقاضی دریافت خدمات بیمه‌ای در طول یک دوره ۱۰ساله که زیر پوشش بیمه قرار می‌گیرد، نیاز به دریافت خسارت دارد یا خیر. هر یک از موارد در فایل نشان‌دهنده مجموعه‌ای از قرار دادهای منعقد شده میان شرکت بیمه و مشتریان است که یکی از آنها خسارت دریافت نموده و باقی قراردادها به دریافت خسارت منتهی نگردیده‌اند. عوامل سن و جنس نیز ثبت گردیده‌اند.
- **Judges.sav** . این فایل شامل داده‌های فرضی در مورد نتایجی است که داوران مجرب و آموزش دیده در مورد عملکرد ۳۰۰ ورزشکار ژیمناستیک اعلام می‌دارند. هر ردیف متعلق به یک ورزشکار بوده و داوران تنها به عملکرد آن ورزشکار توجه نموده‌اند.

- **Kinship-dat.sav** . Rosenberg و Kim در سال ۱۹۷۵، ۱۵ کلمه مربوط به روابط خویشاوندی را مورد مطالعه و تحلیل قرار دارند (عمه، برادر، پسرعمو، دختر، پدر، نوه دختری، پدربزرگ، مادربزرگ، نوه پسری، مادر، پسرخواهر، برادرزاده، خواهر، پسر و عمو). در این مسیر این دو محقق از ۴ گروه از دانشجویان (دو گروه زن و دو گروه مرد) خواستند تا این واژگان را براساس شباهت‌هایشان مرتب نمایند. از دو گروه (یک زن و یک مرد) خواسته شد که این واژگان را دوبار مرتب نمایند به نحوی که در دومین بار، این کار را براساس مقیاس متفاوتی از بار اول انجام می‌پذیرد. بنابراین، در نهایت ۶ منبع به دست می‌آمد. هر کدام از این منابع متشکل از یک ماتریس ۱۵ در ۱۵ بوده که تعداد سلول‌های موجود در آنان برابر با تعداد افرادی است که در آزمایش شرکت کرده‌اند، منهای تعداد دفعاتی که موارد در آن منبع از هم تفکیک شده‌اند.
- **Kinship-ini.sav** . این فایل یک ساختاربندی اولیه از راه‌حلی سه بعدی است برای حل مسئله موجود در فایل Kinship-dat.sav
- **Kinship-var.sav** . این فایل شامل متغیرهای مستقلی همچون جنسیت، نسل و درجه تفکیکی می‌باشد که می‌توانند در تشریح ابعاد راه حل ارائه شده در مسئله مربوط به فایل kinship-dat.sav استفاده گردد. مخصوصاً از این متغیرها می‌توان برای محدود کردن فضای راه‌حل به مجموعه‌ای خطی از این متغیرها استفاده کرد.
- **Mailresponse.sav** . این فایل شامل داده‌های فرضی است در مورد بررسی‌های یک شرکت تولیدی پوشاک در مورد این که استفاده از سیستم پستی پیشتاز و استفاده از ارسال پست الکترونیک به‌صورت مستقیم نتیجه‌ای سریعتر از ارسال پست‌های الکترونیک متعدد جهت تبلیغ محصولات دارد یا خیر؟! افرادی که مسئول دریافت سفارشات مشتریان هستند، مدت زمانی را که پس از فرستادن پست الکترونیک طول کشیده تا مشتری سفارشی را داشته باشد، ثبت می‌نمایند.
- **Marketvalues.sav** . این فایل حاوی داده‌های مربوط به خریدهای خانگی حد فاصل سال‌های ۲۰۰۰-۱۹۹۹ در Algonquin می‌باشد. این اطلاعات جزو اطلاعات عمومی بوده و در دسترس عموم است.

- **Mutualfund.sav** . این فایل حاوی اطلاعاتی در مورد بازار سهام مربوط به سهام شرکت‌های مختلف است. لیست اسامی آنان در S&P500 آورده شده است. هر یک از موارد موجود در فایل متعلق به یک شرکت مجزا است.
- **Nhis2000-subset.sav** . سازمان ملی ممیزی سلامت با استفاده از مصاحبه (NHIS)، سازمان ممیزی بزرگی در آمریکا است که به بررسی سلامت مردم غیرنظامی در این کشور می‌پردازد. مصاحبات به‌صورت حضوری و برروی نمونه‌هایی از خانوارهای آمریکایی انجام می‌گیرد. مشاهدات انجام گرفته برروی رفتارها و عادات افراد موجود در این خانوارها به همراه اطلاعات حاصل از آمارگیری‌های نفوس را برای هر یک از این اعضا به‌دست می‌آورند. این فایل حاوی اطلاعات مربوط به ۲۰۰۰ مورد است.
- **Ozone.sav** . در این فایل اطلاعات مربوط به ۳۳۰ مشاهده و ارزیابی ۶ متغیر هواشناسی که در پیش‌بینی میزان غلظت لایه ازون کاربرد دارند، ثبت شده است. تحقیقات صورت گرفته در این زمینه نشان می‌دهد که رابطه‌ای غیرخطی میان این متغیرها وجود دارد که این مسئله استفاده از رگرسیون استاندارد را با مشکل روبه‌رو می‌نماید.
- **Pain-medication.sav** . این فایل محتوی داده‌هایی فرضی در مورد نتایج حاصل از آموزش روش‌های درمانی ضدحساسیت برای درمان دردهای مفصلی می‌باشد. مسائلی که بیش از همه بررسی آنان مهم می‌باشد این است که چه مدت زمان طول می‌کشد تا دارو تأثیر درمان خود را ایجاد نماید و اینکه چگونه می‌توان روش جدید را با روشی که در گذشته از آن استفاده شده است، مقایسه نمود.
- **Patient-los.sav** . این فایل شامل داده‌های فرضی در مورد سوابق درمانی بیمارانی است که به گمان ابتلا به بیماری قلبی به بیمارستان مراجعه نموده‌اند. هر یک از موارد موجود در فایل به یکی از بیماران تعلق داشته و بسیاری از متغیرهای مرتبط به مدت بستری این بیماران در بیمارستان در آن ثبت گردیده است.
- **Patlos-sample.sav** . این فایل شامل داده‌های فرضی در مورد سوابق درمانی نمونه‌ای از بیماران است که جهت درمان بیماری قلبی عروقی اشان تحت درمان thrombolytics قرار گرفته‌اند. هر یک از موارد موجود در فایل به یکی از بیماران تعلق داشته و اطلاعات مربوط به بسیاری از متغیرهای مرتبط با مدت زمان بستری این بیماران در آن ثبت گردیده است.

- **Polishing.sav** . این فایل شامل داده‌های مربوط به Nambeware polishing Times می‌باشد. این داده‌ها مربوط به تلاش‌های یک تولیدکننده کارد و چنگال برای تنظیم برنامه زمان‌بندی تولید خود می‌باشد. اطلاعات مربوط به هر یک از این محصولات که در فایل ثبت می‌شوند، عبارتند از قطر، مدت زمان پولیش کردن، قیمت و نوع تولید محصول.
- **Poll-cs.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های ناظرین انتخابات برای اندازه‌گیری سطح اشتیاق و پشتیبانی عمومی از یک لایحه قبل از طرح آن در مجلس است. موارد موجود در فایل متعلق به رأی‌دهندگان ثبت شده‌اند. در هر فایل اطلاعاتی در مورد کشور، شهرستان، و همسایگی رأی‌دهندگان ثبت شده است.
- **Property-assess.sav** . این فایل محتوی داده‌های فرضی در مورد تلاش‌های ارزیابان مالی یک کشور است که سعی دارند اطلاعات مربوط به دارایی‌های افراد را با توجه به محدودیت منابع، به روز نگه دارند. موارد موجود در فایل متعلق به ویژگی‌های فروش انجام گرفته در کشور در طی سال گذشته است.
- **Property-assess-cs.sav**. این فایل محتوی داده‌های فرضی در مورد تلاش‌های ارزیابان یک ایالت است که سعی دارند اطلاعات مربوط به دارایی‌های افراد را با توجه به محدودیت منابع به روز نگه دارند. موارد موجود در فایل محتوی دارایی‌های موجود در ایالت است، در این فایل اطلاعات مربوط به کشور، استان، مناطق مجاوری ثبت شده است که سرمایه‌ها در آن قرار دارند، مدت زمانی که از آخرین ارزیابی گذشته است و میزان دارایی که در آن زمان ارزیابی گردیده است، ثبت شده و قرار دارد.
- **Property-assess-cs-sample.sav** . این فایل شامل داده‌هایی فرضی از یک نمونه از دارایی‌هایی است که در فایل Praperty-assess-cs.sav آمده است. نمونه براساس طراحی خاص در فایل property-assess-csplan برداشته شده است و در این فایل احتمالات و وزن‌دهی نمونه را ثبت می‌نمایند. همچنین متغیر current value نیز پس از این که نمونه انتخاب گردید به متغیرهای موجود اضافه می‌گردد.
- **Recidivism.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های یک مؤسسه دولتی اجرای قانون برای درک بهتری از میزان تکرار جنایات در حوزه قضایی مربوط به آن مؤسسه است، هر یک از موارد موجود در فایل متعلق به یک خلافکار قدیمی بوده و اطلاعاتی همچون آمارگیری نفوس، برخی جزئیات مربوط به جرائم اولیه آنان و همچنین

زمانی که تا دستگیری دوم طول کشیده است. (چنانچه جرم دوم تا دوسال بعد از دستگیری اول روی دهد) در فایل و برای آن خلافکار ثبت گردیده است.

- **Recidivism-cs-sample.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های انجام گرفته از سوی یک مؤسسه دولتی اجرای قانون برای درک بهتری از میزان تکرار جنایات در حوزه قضایی مربوط به آن مؤسسه است. هر مورد موجود در فایل متعلق به یکی از خلافکار قدیمی است که بعد از دستگیری اول خود در طول ماه ژوئن ۲۰۰۳ آزاد شده است و اطلاعاتی همچون آمارگیری نفوس، برخی جزئیات مربوط به جرم اول او و همچنین اطلاعات مربوط به دستگیری دوم او چنانچه این دستگیری تا پایان ماه ژوئن ۲۰۰۶ ، روی دهد نیز در فایل او ثبت گردیده است. متخلفینی که در طی برنامه نمونه‌گیری انتخاب شده‌اند در فایل recidivism-cs.csplan ثبت گردیده‌اند.
- **Rfm-transactions.sav** . این فایل شامل داده‌های فرضی در مورد فرایند خرید از جمله روز انجام خرید، وسیله و موردی که خریده شده است و میزان پولی که در هر خرید پرداخت شده است، می‌باشد.
- **Salesperformance.sav** . این فایل شامل داده‌های فرضی در مورد ارزیابی ۲ دوره آموزشی جدید که برای فروشندگان برگزار می‌گردد، می‌باشد. ۶۰ نفر از کارمندان را در نظر گرفته، سپس آنان را در سه گروه تقسیم نمودند و تمامی آنان را مورد آموزش‌های استاندارد قرار دارند. علاوه بر این آموزش استاندارد، در مورد گروه ۲ آموزش تکنیکی و در مورد گروه ۳ آموزش نحوه استفاده بهتر از فرایندهای دستی (کارهایی که با دست انجام می‌گیرند) اعمال گردید. هر یک از کارمندان را پس از پایان دوره مورد آزمون قرار داده و نتیجه آزمون را ثبت نمودند، هر یک از موردهای موجود در فایل متعلق به یکی از دوره دیده‌ها بوده و نتایج حاصل از آزمون انجام شده در مورد آن شخص نیز در فایل مربوط به آن فرد، ثبت گردیده است.
- **Satisf.sav** . این فایل شامل داده‌هایی در مورد ارزیابی انجام شده توسط یک شرکت خرده فروشی در ۴ محل است. در کل ۵۸۲ نفر از مشتریان مورد ارزیابی قرار گرفته و هر مورد شامل عکس‌العمل‌ها و نتایجی است که از هر یک از مشتریان به دست آمده است.
- **sav** . این فایل شامل اطلاعاتی در مورد ویژگی‌ها و خصوصیات پیچ‌گوشتی‌ها، پیچ و مهره‌ها، چرخ‌دنده‌ها و پونزها است.

- **Shampoo-ph.sav** . این فایل شامل داده‌های فرضی در مورد فرایند کنترل کیفیت یک کارخانه تولید محصولات مربوط به مو می‌باشد. در بازه‌های زمان مشخص، ۶ دسته خروجی مجزا از محصولات مورد بررسی قرار می‌گیرند و میزان PH آنان ثبت می‌گردد. میزان دامنه موردنظر و مناسب این مؤلفه ۵/۵-۴/۵ می‌باشد.
- **Ships.sav** . مجموعه داده‌ای که توسط Mccullagh در سال ۱۹۸۹ مورد آنالیز قرار گرفته و ارائه گردیده است. در مورد آسیب‌هایی است که به واسطه موج‌های دریا به کشتی‌های باربری وارد می‌آید. این مجموعه داده در این فایل آمده است. میزان شیوع این آسیب‌ها را می‌توان با نسبت Poisson محاسبه نمود. برای این منظور می‌بایست مواردی همچون نوع کشتی، زمانی که تولید شده است و دوره‌هایی که در آن تعمیر می‌گردد مشخص باشد.
- **Site.sav** . این فایل شامل داده‌هایی فرضی در مورد تلاش‌های یک شرکت برای یافتن مکان‌های جدید برای رشد و پیشرفت تجارتشان است. این شرکت دو مشاور را استخدام نموده است که به‌صورت مجزا، به بررسی این مکان‌ها بپردازند و در کنار مشروح گزارش خلاصه‌ای از آن را به‌صورت "خوب" ، "نسبتاً خوب" و یا "نامناسب" ارائه دهند.
- **Siteratings.sav** . این فایل شامل داده‌هایی فرضی در مورد آزمون اشعه بتا در وب سایت جدید یک شرکت بازرگانی الکترونیک، می‌باشد. هر یک از موارد موجود در فایل متعلق به یکی از آزمون‌های اشعه بتا است که قابلیت استفاده از سایت را در مقیاسی از ۲۰ -۰ مشخص می‌نماید.
- **Smokers.sav** . این فایل خلاصه‌ای از بررسی‌های سازمان ملی خانوار امریکا در سال ۱۹۹۸ است که در مورد سوء مصرف داروها، صورت گرفته است و می‌تواند نمونه‌ای احتمالی از خانوارهای امریکایی قلمداد گردد. بنابراین اولین قدم در تحلیل این داده‌ها، می‌بایست وزن‌دهی داده‌ها باشد.
- **Smoking.sav** . این فایل شامل داده‌هایی فرضی است که توسط Greenacre در سال ۱۹۸۴ ارائه گردیده است. جدول علاقه‌مندی‌ها برای سیگاری‌ها با توجه به دسته‌بندی کاری آنها شکل گرفته است. متغیر Staff group شامل دسته‌بندی‌های شغلی زیر است:
- Sr managers، Jr manager، Sr employees، Jr employees، Secretaries، همچنین دسته دیگری تحت عنوان National Average نیز در این دسته‌بندی وجود دارد که گاهاً به‌عنوان دسته‌ای تکمیلی در آنالیزها آورده می‌شود. متغیر Smoking شامل رفتارهای زیر می‌شود:

- None، Light، Medium، Heavy، همچنین دسته‌بندی‌هایی همچون No Alcohol و Alcohol نیز وجود دارند که در آنالیزها به‌صورت دسته‌بندی‌های تکمیلی آورده می‌شوند.
- **Storebrand.sav** . این فایل شامل داده‌های فرضی است در مورد تلاش‌های صورت گرفته از سوی مدیر یک فروشگاه خواربار فروشی که تلاش دارد میزان فروش مواد پاک‌کننده‌ای که در انبار باقی مانده‌اند را در مقابل نام‌های تجاری دیگر، افزایش دهد. او تصمیم گرفت برروی این محصولات تبلیغ بیشتری داشته و به هنگام خرید با مشتریان در مورد این نام‌های تجاری صحبت نموده و آنان را به خرید این محصولات ترغیب نماید. هر مورد وجود در این فایل متعلق به یک مشتری خاص است.
- **Stores.sav** . این فایل شامل داده‌های فرضی در مورد میزان فروش در بازار برای دو مغازه خواروبار فروشی است که با هم در رقابت هستند. هر مورد موجود در فایل، متعلق به میزان اشتراک از بازار در یک ماه مشخص است.
- **Stroke-invalid.sav** . این فایل شامل داده‌های فرضی در مورد مقادیر اولیه موجود در پایگاه داده درمانی است و شامل خطاهای بسیاری در زمان وارد کردن داده‌ها است.
- **Stroke-survival** . این فایل شامل داده‌های فرضی در مورد طول مدت حیات بیمارانی است که از یک برنامه توانبخشی خارج می‌شوند.
- **Survey-sample.sav** . این فایل شامل داده‌های فرضی در مورد داده‌های بررسی شده‌ای همچون، آمارگیری نفوس و انواع مختلف محاسبات انجام گرفته برروی آنان می‌باشد.
- **Tastetest.sav** . این فایل شامل داده‌های فرضی در مورد تأثیر رنگ کودگیاهی برروی طعم محصول است. توت‌فرنگی در کودهایی به رنگ قرمز، آبی، و سیاه رشد می‌کند که با توجه به مقیاس ۱ تا ۵ درجه‌بندی می‌شوند. (بسیار پایین میانگین تا بسیار بالای میانگین). هر یک از موارد موجود در فایل متعلق به یک آزمون مزه مشخص است.
- **Telco.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های یک شرکت ارتباط از راه‌دور که سعی دارد میزان نوسان مشتریان خود را کاهش دهد. هر یک از موارد موجود در فایل متعلق به یکی از مشتریان بوده و اطلاعات متنوعی از خدمات مورد استفاده آن مشتری و اطلاعات آماری مربوط به او در این فایل ثبت شده است.
- **Telco-missing.sav** . این فایل زیرمجموعه‌ای از فایل telco.sav است که در آن برخی از مقادیر مربوط به آمارگیری نفوس با مقادیر جایگزین دیگری، جابه‌جا شده‌اند.

- **Testmarket.sav** . این فایل شامل داده‌های فرضی در مورد برنامه یک اغذیه فروشی برای اضافه نمودن یک نمونه جدید به لیست غذاهای خود است. سه دستورالعمل ممکن برای تبلیغ غذای جدید در نظر گرفته شده است. بدین‌ترتیب این غذا به‌صورت تصادفی در بسیاری از نمایندگی‌ها معرفی می‌گردد. در هر یک از نمایندگی‌ها شیوه‌ای متفاوت برای تبلیغ غذای جدید استفاده می‌گردد و میزان فروش هفتگی در طی ۴ هفته ابتدایی ثبت می‌گردد. هر یک از موارد موجود در فایل به اطلاعات ثبت شده هر یک از نمایندگی‌های مختلف و میزان فروش هفتگی آنان تعلق دارد.
- **Viruse.sav** . این فایل شامل داده‌های فرضی در مورد تلاش‌های صورت گرفته از سوی یک شرکت تأمین‌کننده خدمات اینترنتی است که سعی دارد میزان تأثیر ویروس‌ها برروی شبکه خود را محاسبه نماید. این شرکت با میزانی از ترافیک پست‌های الکترونیک حاصل از فعالیت ویروس‌ها در شبکه خود روبه‌رو است.
- **Wheeze.steubenville** . این فایل شامل نتایج حاصل از آزمونی مستمر برروی تأثیراتی است که آلودگی هوا برروی سلامت کودکان می‌گذارد. داده‌ها شامل نتایج اندازه‌گیری‌های متعدد صورت گرفته در مورد وضعیت تنفس کودکان ۷ ، ۸ ، ۹ و ۱۰ ساله در Ohio امریکا است. بدین‌ترتیب که آیا نفس کشیدن این کودکان همراه با صدای خس‌خس می‌باشد یا خیر. در ضمن این که آیا مادر در طی سال اول انجام مطالعه سیگار مصرف می‌نمود یا خیر نیز در فایل ثبت شده است.
- **Workprog.sav** . این فایل شامل داده‌های فرضی در مورد برنامه کاری دولت می‌باشد که در آن سعی بر این دارد که افراد کم بازده را در شغل مناسب‌تری قرار دهد. نمونه‌ای از متقاضیان بالقوه این برنامه را تحت نظر گرفته و تعدادی از آنان را به‌صورت تصادفی انتخاب نموده‌ایم. هر مورد موجود در فایل متعلق به یکی از شرکت‌کنندگان در ارزیابی است.